Reinforcement learning in games

By

David S. Leslie

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science

February 2004

School of Mathematics

PAGE NUMBERING AS ORIGINAL

Abstract

This thesis investigates the properties of learning in games, where the information available to each player does not include a specification of the game or observations of opponent play. Each player responds only to the rewards they received for playing particular actions on previous plays of the game.

We relate two modifications of Börgers and Sarin's stimulus-response learning (Börgers and Sarin 1997) to the replicator dynamics. In these algorithms, observed rewards are used to directly modify the strategies of the players.

Then an example of actor–critic learning, in which a value function is used to adapt the strategies, is studied using two-timescales stochastic approximation to show that the strategies track the smooth best response dynamics. An extension, in which players learn at different rates, is analysed using a newly-developed theory of multiple-timescales stochastic approximation (Leslie and Collins 2003).

Q-learning in games, where the strategies are simply functions of value estimates, is then studied using similar methods, employing smooth best responses and player-dependent learning rates.

A modified actor–critic algorithm is introduced, in which strategies adapt towards a best response (instead of a smooth best response). This is analysed, by generalising some results on fictitious play, and shown to converge in several classes of games.

Initial investigations into extending these algorithms to stochastic games study the contraction properties of classes of smooth best responses.

Acknowledgements

This thesis is dedicated to my parents, who taught me how to think.

Many thanks are due to my principal adviser, Sean Collins, for his hard work, assistance, and general support throughout my studies, and to Andy Wright, for identical reasons. Thanks also to the entire Department of Mathematics at the University of Bristol, for providing an inspiring and supportive place to study.

My research has been supported by CASE Research Studentship 00317214 from the UK Engineering and Physical Sciences Research Council in cooperation with BAE SYSTEMS.

Declaration

I, the author, declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

The views expressed in the dissertation are those of the author and in no way represent those of the University of Bristol.

The dissertation has not been presented to any other University for examination either in the United Kingdom or overseas.

David S. Leslie

31/3/04

Date

Contents

List of Figures

Chapter 1

Introduction and literature review

Our areas of study are games and Markov decision process. These fields were initiated by Von Neumann, Morgenstern, Nash, Bellman, Shapley and others in the 1940s and 50s, and have been studied extensively since. The basic premises of these areas are that a player chooses actions and gains rewards in return, with each player attempting to maximise their own reward.

1.1 Normal form games

Game theory arose from the study of conflict, where a player's rewards depend not only on their own actions but also on the actions of others. The initial inspiration was the Cold War, but applications are to be found in economics, evolution, machine learning and many other areas. The formal definition of a normal form game is:

- A finite set of players $(1, \ldots, N)$,
- A finite set of actions A^i for each player $i = 1, \ldots, N$, resulting in a finite set of joint actions $\underline{A} = A^1 \times \cdots \times A^N$,
- A reward function $r^i : \underline{A} \to \mathbb{R}$ for each player $i = 1, \ldots, N$, where $r^i(\underline{a})$ is the reward given to player i if joint action $\underline{a} \in \underline{A}$ is played.

This is a modern formulation of the framework first proposed by Von Neumann and Morgenstern (1953), which in turn brings together previous work, largely by Von Neumann.

There are several special classes of games, for which the theory is more developed in some areas. Two of these are detailed here.

Zero (or constant) sum games: $\sum_i r^i(\underline{a}) = 0$ (or more generally a constant) for any $\underline{a} \in \underline{A}$. These games can be considered as entirely competitive, since one player's gain is another's loss.

Partnership games: $r^i(\underline{a}) = r^j(\underline{a})$ for all i, j and joint actions $\underline{a} \in \underline{A}$. This is the opposite end of the spectrum to zero sum games, since all players get identical rewards, and so face a joint maximisation problem.

Occasionally we will use the term "$n \times m$ game" for a game where player 1 (resp. 2) has n (resp. m) actions.

1.1.1 Classical game theory

A player adopts a strategy to play the game, which is a rule telling them which action to pick. Nash (1950) defines a (Nash) equilibrium of a game to be a strategy for each player such that no player can increase their reward by unilaterally deviating from their strategy. If we restrict these strategies to choosing single actions then not all games will admit such an equilibrium.

Let Δ^i denote the set of probability distributions over the action space A^i of player i. A mixed strategy for player i is an element $\pi^i \in \Delta^i$; henceforth we will call a strategy that chooses a single action with probability 1 a pure strategy. Defining $\Delta = \Delta^1 \times \cdots \times \Delta^N$, we say that a joint strategy is an element $\pi = (\pi^1, \ldots, \pi^N) \in \Delta$. Note that it is implicitly assumed that players implement their strategies independently. There are unique multilinear extensions of the reward functions to this mixed strategy space, and in standard abuse of notation we will

denote by $r^i(\pi)$ the expected reward to player i when the players use joint mixed strategy π. Similarly we define $r^i(a^i, \pi^{-i})$ to be the expected reward to player i when they play action a^i and all players other than i play according to the opponent strategy $\pi^{-i} = (\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^N)$.

Given an opponent strategy π^{-i}, player i has a set of best responses

$$\mathrm{BR}^i(\pi^{-i}) = \{b^i \in \Delta^i : r^i(b^i, \pi^{-i}) = \max_{\pi' \in \Delta^i} r^i(\pi', \pi^{-i})\}; \tag{1.1}$$

$\mathrm{BR}^i(\pi^{-i})$ is the set of mixed strategies which maximise player i's expected reward, given that opponents play strategy π^{-i}. The best response correspondence is defined as

$$\mathrm{BR}(\pi) = \{(b^1, \ldots, b^N) \in \Delta : b^i \in \mathrm{BR}^i(\pi^{-i}) \text{ for each } i\}, \tag{1.2}$$

where π^{-i} are the opponent strategies arising from joint strategy π.

A Nash equilibrium $\tilde{\pi} \in \Delta$ is therefore a fixed point of the best response correspondence BR:

$$\tilde{\pi} \in \mathrm{BR}(\tilde{\pi}). \tag{1.3}$$

Theorem 1 (Nash 1950) *Every game has at least one equilibrium.*

Since $\tilde{\pi}^i \in \mathrm{BR}^i(\tilde{\pi}^{-i})$ for all i at a Nash equilibrium $\tilde{\pi}$, we see that

$$r^i(\tilde{\pi}) \geq r^i(\pi^i, \tilde{\pi}^{-i}) \quad \text{for all } \pi^i \in \Delta^i, \tag{1.4}$$

and so at a Nash equilibrium, no player can improve their expected reward by a unilateral deviation away from $\tilde{\pi}^i$. If, for each i, the inequality (1.4) is strict for all $\pi^i \neq \tilde{\pi}^i$, $\tilde{\pi}$ is called a strict Nash equilibrium. Note that in this case, π^i must be a pure strategy for each i; supposing $\tilde{\pi}$ is a mixed strategy Nash equilibrium, with $\tilde{\pi}^i(a) > 0$ and $\tilde{\pi}^i(b) > 0$ for $a, b \in A^i$, then we must have $r^i(a, \tilde{\pi}^{-i}) = r^i(b, \tilde{\pi}^{-i})$ or (1.4) would not hold, and so constructing the strategy π^i by taking $\pi^i(a) = \tilde{\pi}^i(a) + \tilde{\pi}^i(b)$, $\pi^i(b) = 0$, and $\pi^i(a') = \tilde{\pi}^i(a')$ otherwise gives $r^i(\pi^i, \tilde{\pi}^{-i}) = r^i(\tilde{\pi})$, so $\tilde{\pi}$ is not a strict Nash equilibrium.

Chapter 1. Introduction and literature review

For two player zero sum games there is a particular solution concept, known as the maximin solution. Here players play to maximise the minimum possible reward they might get.

Theorem 2 (Von Neumann 1928) *For a 2-player zero-sum game let*

$$E^1 = \left\{ a^1 \in A^1 : \min_{a^2 \in A^2} r^1(a^1, a^2) = \max_{b^1 \in A^1} \min_{a^2 \in A^2} r^1(b^1, a^2) \right\}$$

and similarly for E^2. If player i plays a (possibly mixed) strategy consisting only of actions in E^i, then an equilibrium will result.

However in general games, where players are not in direct competition, it does not make sense for a player to take this conservative approach of maximising the worst possible reward they might receive, since it may be the case that the other players will not gain by minimising the reward available to the player.

Another concept of classical game theory is that of dominance (Nash 1951): we say a pure strategy a^i is strictly dominated by pure strategy b^i if $r^i(a^i, \pi^{-i}) < r^i(b^i, \pi^{-i})$ for all opponent mixed strategies π^{-i}. A rational player will clearly never play a strictly dominated strategy, no matter what the other players are doing. It therefore makes sense to consider the game with a dominated strategy removed. In this new game there may well be further strategies that are strictly dominated which can now be removed. This process of iterative elimination of strictly dominated strategies does not alter the Nash equilibria of the game—no Nash equilibrium will contain a strategy that can be eliminated by iterative strict dominance.

Although this classical theory shows that at least one equilibrium must exist in all games, and may help to calculate these equilibria, there are many philosophical problems remaining:

- We have observed that all actions played with positive probability at a Nash equilibrium will receive the same expected reward. There is therefore no immediate incentive for any particular player to play the Nash equilibrium strategy instead of a different strategy using the same set of actions.

- Most games have more than one equilibria. Therefore players must coordinate to play a particular equilibrium.

- In applications of game theory in economics and biology, it is often observed that Nash equilibrium strategies are played, even though the 'players' do not know that they are playing a game, let alone know the reward functions and calculate the Nash equilibrium.

While the first two points have been addressed, at least in theory, by the work of Harsanyi and Selten (Harsanyi 1973; Harsanyi and Selten 1988) the last remains an open problem, being considered at length in recent textbooks on learning theory (Fudenberg and Levine 1998) and evolutionary game theory (Hofbauer and Sigmund 1998).

1.1.2 Smooth best responses and Nash distributions

Harsanyi (1973) showed that approximately correct equilibrium strategies arise naturally if players are unsure about the payoff functions defining the game. For this thesis, we note from Hofbauer and Sandholm (2002) that, under a simplifying assumption, this is equivalent to players choosing a smooth best response

$$\beta^i(\pi^{-i}) = \underset{\pi^i \in \Delta^i}{\operatorname{argmax}} \left\{ r^i(\pi^i, \pi^{-i}) + \tau v^i(\pi^i) \right\} \tag{1.5}$$

to opponent strategies π^{-i}, where $\tau > 0$ is a temperature parameter and $v^i : \Delta^i \to \mathbb{R}$ is player i's smoothing function, which is a smooth, strictly differentiably concave function such that as π^i approaches the boundary of Δ^i the slope of v^i becomes infinite (Fudenberg and Levine 1998, Chapter 4). The conditions on v^i imply that β^i is a well-defined, continuous function (in contrast with BR^i defined in (1.1) which is a discontinuous, possibly multiple-valued correspondence). Further, since the slope of v^i becomes infinite as π^i approaches the boundary of Δ^i, the strategy $\beta^i(\pi^{-i})$ is a completely mixed strategy, i.e. $\beta^i(\pi^{-i})(a^i) > 0$ for each a^i.

If players choose smooth best responses (1.5) instead of best responses (1.1), equilibrium play will deviate from Nash equilibrium. An easy illustration of this is if a game has only pure strategy equilibria: since all actions are played with positive probability under a smooth best response, no player will ever play a Nash equilibrium strategy. Instead we define a Nash distribution in an analogous way to Nash equilibria, but using smooth best responses instead of best responses. Writing $\beta = (\beta^1, \ldots, \beta^N)$, a Nash distribution $\tilde{\pi}$ satisfies

$$\tilde{\pi} = \beta(\tilde{\pi}).$$

It is clear, from the Brouwer fixed point theorem, that at least one Nash distribution exists for any game. However, the Nash distributions of a game are not properties only of the game, but also of the smooth best response functions β^i (which depend on the choices of τ and v^i).

Although Harsanyi's original purification theorem does not quite apply in the situation we describe here (it relies on the fact that each player's prior distributions over the rewards has finite support) Govindan *et al.* (2003) show that for sufficiently small temperatures τ there is a Nash distribution close to any Nash equilibrium.

1.1.3 Fictitious Play

The first attempt to explain how Nash equilibrium might be achieved other than by an introspective study of the game was by Brown (1951). Although originally proposed as a computational method to calculate the equilibrium of a game, Brown's process can also be considered as a learning procedure (Fudenberg and Levine 1998). This process, known as fictitious play, assumes that a game is played repeatedly. Each player stores the frequency with which opponents have played each of their actions in the past. When the game is played, each player plays a best response to these empirical frequency vectors.

Let the game be played at time steps $n = 1, 2, \ldots$ Just before the nth game is

played, the empirical frequency vector of player i's past play is given by σ_n^i (with σ_0^i chosen arbitrarily); write $\sigma_n = (\sigma_n^1, \ldots, \sigma_n^N)$ and $\sigma_n^{-i} = (\sigma_n^1, \ldots, \sigma_n^{i-1}, \sigma_n^{i+1} \ldots, \sigma_n^N)$. Player i then chooses an action $b_n^i \in \mathrm{BR}^i(\sigma_n^{-i})$. Writing $b_n = (b_n^1, \ldots, b_n^N)$ we see that

$$\sigma_{n+1} = \left(1 - \frac{1}{n+1}\right)\sigma_n + \frac{1}{n+1} b_n \quad \text{where} \quad b_n \in \mathrm{BR}(\sigma_n). \tag{1.6}$$

At each time step n, the players will (generically) choose a pure strategy best response, and the strategies used by the players cannot converge to a mixed equilibrium. However the vector of empirical frequencies σ_n can converge to such an equilibrium. Indeed we have the following:

Theorem 3 (Fudenberg and Kreps 1993) *Suppose that in a fictitious play process the empirical frequencies σ_n converge to a point $\sigma \in \Delta$ as $n \to \infty$. Then σ is a Nash equilibrium.*

We say a game has the fictitious play property if the empirical frequencies σ_n of any fictitious play process in that game will converge to a Nash equilibrium for any initial conditions σ_0.

Theorem 4 *Games in the following classes have the fictitious play property:*

1. *2-player zero-sum games (Robinson 1951),*
2. *N-player partnership games (Monderer and Shapley 1996),*
3. *games solvable by iterated strict dominance (Milgrom and Roberts 1991), and*
4. *non-degenerate $2 \times m$ games (Berger 2003).*

However not all games have the fictitious play property: Shapley (1964) constructed an example of a 2-player game for which the empirical frequencies of a fictitious play process need not converge. Further, Fudenberg and Kreps (1993) demonstrate a 2×2 game which has the fictitious play property, but for which average payoffs do not converge to the payoffs of the Nash equilibrium.

To combat some of these problems, stochastic fictitious play was introduced by Fudenberg and Kreps (1993), and studied further by Benaïm and Hirsch (1999). Here it is (essentially) assumed that players play a smooth best response (1.5) to the empirical frequency vector σ_n, for some choice of temperature parameter τ and smoothing functions v^i that is fixed throughout the learning process. This allows convergence of players' actual strategies (as opposed to the empirical frequencies) and hence removes many of the subtle problems related to convergence of (traditional) fictitious play. The analysis of stochastic fictitious play requires the use of stochastic approximation theory, and will be addressed in Chapter 3.

A version of best-response adaptation for fully rational Bayesian learners is given by Kalai and Lehrer (1993a, 1993b). In their framework, a strategy determines play for the entire repeated game, and players attempt to maximise their discounted future rewards. Players initialise by choosing prior distributions for opponent strategies, and these are updated in a Bayesian fashion as actions are observed. Players utilise a best response to their beliefs about future opponent play (as with fictitious play), though in the context of an infinitely repeated game with discounted rewards the calculation of this best response is far from trivial. They show that for general games players will eventually choose actions as if they were following a Nash equilibrium of the discounted repeated game (under continuity conditions on the priors).

1.1.4 Consistent learning

Returning to the context of repeated normal form games, a learning algorithm is defined to be consistent if, asymptotically, the average reward received is as large as the reward that is obtained by the best response to the empirical distribution of opponent play (Hannan 1957). Hart and Mas-Colell (2000) show that the empirical frequency distributions arising if all players follow a consistent learning procedure will converge to the set of correlated Nash equilibria of a game (a

correlated Nash equilibrium is a Nash equilibrium under the influence of some correlating device—we will not study these in this thesis and so simply refer to Fudenberg and Tirole (1991) for the definition).

Fudenberg and Levine (1999) show that stochastic fictitious play is ϵ-consistent (i.e. the average reward received is within ϵ of the reward obtained by the best response to the empirical distribution of opponent play) for suitable choice of temperature parameters, but Hart and Mas-Colell (2001a) show that any algorithm where the strategy updates do not depend on the total reward received so far cannot be consistent. Instead, Hart and Mas-Colell (2001a) present a full class of consistent regret-based strategies, extending previous work (Foster and Vohra 1997, 1998, 1999; Hart and Mas-Colell 2000).

However, the convergence of these consistent algorithms is in the same sense as fictitious play—the empirical distributions of play converge, as opposed to the strategies of the players. Moreover the set of correlated equilibria of a game contains at least the convex hull of the Nash equilibria, and in general it is difficult to know how big the set of correlated equilibria will be (Fudenberg and Tirole 1991).

These consistent algorithms require full information about rewards, i.e. the players need to know what reward each of their actions would have received at each time step, and so are generally not applicable if the game is unknown or if players cannot observe opponent actions. To counter this problem, consistent reinforcement learning algorithms have been developed that can be used in the case of incomplete information. Baños (1968) and Megiddo (1980) construct explicit sequences of exploration and exploitation (see Section 1.2.3) allowing players to learn about the game being played while still playing sufficiently optimally to receive high average reward. These algorithms are shown to asymptotically achieve the minimax payoff in zero-sum games. Auer *et al.* (1995) and Hart and Mas-Colell (2001b) present algorithms where the strategies evolve in a Markovian nature, and still result in consistent behaviour. Thus the empirical frequencies of play converge to the set of correlated equilibria under these reinforcement learning algorithms. We

will attempt to improve on these results in this thesis.

1.1.5 Evolutionary game theory

Maynard Smith (1982) introduced the study of evolutionary processes using game theory; recent developments are summarised by Hofbauer and Sigmund (1998). In this model the role of a player in a game is taken by a very large (infinite) population of individuals, each of which plays a pure strategy $a \in A$. Thus a mixed strategy corresponds to a particular population state $\pi \in \Delta$, where Δ is the set of probability distributions over A. Usually in this setting we consider symmetric games, where the population is essentially playing against itself; a player who uses action a in a population in state $\pi \in \Delta$ will receive reward

$$r(a, \pi) = (U\pi)_a$$

where U is an $|A| \times |A|$ payoff matrix. A Nash equilibrium in this context is a population state $\tilde{\pi}$ such that

$$\tilde{\pi}^T U \tilde{\pi} \geq \pi^T U \tilde{\pi} \quad \text{for all } \pi \in \Delta.$$

A central concept of symmetric games is that of the evolutionarily stable strategy (ESS). This is a strategy such that if a small population of mutants try to invade then they will not be able to gain a foothold by evolutionary means (i.e. they will receive a smaller reward than the individuals of the resident population). A strategy $\tilde{\pi}$ is an ESS if

$$r(\tilde{\pi}, (1-\epsilon)\tilde{\pi} + \epsilon\pi) > r(\pi, (1-\epsilon)\tilde{\pi} + \epsilon\pi) \quad \text{for any } \pi \text{ and for sufficiently small } \epsilon.$$

By the linearity of r in the second argument, we see that this is equivalent to

$$r(\tilde{\pi}, \tilde{\pi}) \geq r(\pi, \tilde{\pi}) \quad \text{for any } \pi, \text{ and in addition}$$

$$\text{if} \quad \pi \neq \tilde{\pi} \quad \text{and} \quad r(\tilde{\pi}, \tilde{\pi}) = r(\pi, \tilde{\pi}) \quad \text{then} \quad r(\tilde{\pi}, \pi) > r(\pi, \pi).$$

Thus an ESS is a Nash equilibrium, but with an additional constraint. Not all games have an ESS, but if an evolutionary game has an ESS then this is a natural fixed point for any evolutionary dynamic.

A fictitious play procedure for symmetric games is given by Hofbauer (1995). Here we start with a population consisting of a single player, then at each stage a new player is added; the new player uses a pure strategy which is a best response to the current population state. Thus

$$\pi_{n+1} = \left(1 - \frac{1}{n+1}\right)\pi_n + \frac{1}{n+1}b_n \quad \text{where} \quad b_n \in \mathrm{BR}(\pi_n).$$

Hofbauer (1995) shows that this process will converge to the set of "segregation equilibria" of the game (a concept used only within that paper). For zero-sum games with an interior equilibrium, and games with an interior evolutionarily stable strategy, Hofbauer (1995) then shows that this set coincides with the set of Nash equilibria. A stochastic version of this process, analogous to the stochastic fictitious play of Fudenberg and Kreps (1993), is introduced and studied by Hofbauer and Sandholm (2002).

1.1.6 Dynamical systems in game theory

Many previous studies of adaptation in games have considered 'small time-step limits' and/or 'infinite population limits', and investigate the resultant dynamical systems. We present three differential equations arising from game-theoretical ideas; convergence results for these systems will be given in the relevant chapters. More general dynamics have been studied, for example by Hopkins (1999). Although we don't need to study these here, it is interesting to note (Gaunersdorfer and Hofbauer 1995; Hopkins 1999) that there are relations between the three dynamics we present, even though the founding principles of the (smooth) best response dynamics and the replicator dynamics are very different.

Best response (BR) dynamics

These dynamics can be considered as a continuous version of fictitious play (1.6) (Brown 1951; Hofbauer 1995), and have been studied recently by Gilboa and Matsui (1991) and Gaunersdorfer and Hofbauer (1995). Hofbauer (1995) concentrates

on the symmetric game case, which (due to our choices of notation) can be specified in exactly the same way. The idea is that strategies (or population states) will adjust towards a best response to the current strategy (population state). This results in the dynamics

$$\dot{\pi} \in \mathrm{BR}(\pi) - \pi, \tag{1.7}$$

which actually defines a differential inclusion (Aubin and Cellina 1984), since $\mathrm{BR}(\pi)$ is not single-valued for all π. Fixed points of this dynamical system are clearly Nash equilibria, so if a trajectory converges to a point $\tilde{\pi}$ then $\tilde{\pi}$ must be a Nash equilibrium. On the other hand, Cowan (1992) shows that chaotic behaviour can arise from the BR dynamics.

Smooth BR dynamics

This is a simple modification of the BR dynamics, where instead of adapting towards a (possibly not completely specified) best response, players use a smooth best response function β^i instead, resulting in the dynamics

$$\dot{\pi} = \beta(\pi) - \pi. \tag{1.8}$$

This has been considered in the N-player case by Hofbauer and Hopkins (2000), and in the symmetric games case by Hofbauer and Sandholm (2002). Fixed points of this dynamic are not Nash equilibria, but instead are Nash distributions. However, as noted previously, for small temperature parameters τ in the definition (1.5) of the smooth best response, these Nash distributions will be close to the equilibria of the game.

Replicator dynamics

These dynamics are central to evolutionary game theory, but are also of interest when studying learning (Börgers and Sarin 1997). In the evolutionary setting they arise from the population view of mixed strategies, where the number of players playing a pure strategy will grow at a rate proportional to the reward received by

that pure strategy against the mixed strategies of the opponent populations. Consider a population in player role i: if we let $n^i(a^i)$ be the number of players playing pure strategy a^i and $n^i = \sum_{a^i \in A^i} n^i(a^i)$ be the number of players in population i then $\dot{n}^i(a^i) = r^i(a^i, \pi^{-i})n^i(a^i)$. Letting $\pi^i(a^i) = \frac{n^i(a^i)}{n^i}$, we see that

$$\begin{aligned}\dot{\pi}^i(a^i) &= \frac{\dot{n}^i(a^i)}{n^i} - \frac{n^i(a^i)\dot{n}^i}{(n^i)^2} \\ &= r^i(a^i, \pi^{-i})\pi^i(a^i) - \pi^i(a^i) \sum_{b^i \in A^i} r^i(b^i, \pi^{-i})\pi^i(b^i) \qquad (1.9) \\ &= \pi^i(a^i)\left(r^i(a^i, \pi^{-i}) - r^i(\pi)\right)\end{aligned}$$

This is the asymmetric replicator dynamics (the symmetric replicator dynamics, used in evolutionary biology, is given by the same equation but with all superscripts removed). There is also a version of the asymmetric replicator dynamics known as the payoff-normalised replicator dynamics. These are only well-defined if the payoffs are positive, since otherwise division by zero could occur, and have the form

$$\dot{\pi}^i(a^i) = \pi^i(a^i)\frac{r^i(a^i, \pi^{-i}) - r^i(\pi)}{r^i(\pi)}. \qquad (1.10)$$

Maynard Smith (1982) states that there is "room for doubt as to which form is more appropriate". Fixed points for all versions of the replicator dynamics are Nash equilibria for the game.

Non-convergence of dynamical systems

In a forthcoming paper, Hart and Mas-Colell (2003) use a simple example to show that no "uncoupled" dynamics can lead to Nash equilibrium in general normal-form games. An uncoupled dynamic is any dynamic where the evolution of strategy π^i depends only on r^i and π, and therefore includes all three dynamics considered above (although clearly the smooth BR dynamics (1.8) cannot converge to Nash equilibrium in general, since fixed points of the system are Nash distributions). The property of being uncoupled also holds for virtually all dynamical systems studied in game theory. This result shows that any system in which players act

naively and ignore the payoffs of the others cannot converge to equilibrium in general games.

1.2 Reinforcement learning

Many of the learning processes previously studied in game theory require the players to observe opponent play and make calculations, possibly based on knowledge of the reward functions. In an attempt to consider systems where players do not know the structure of the game, or even that they are playing a game, we would like players to learn the value of their actions simply from experience of playing those actions. Reinforcement learning (Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998) provides a method for doing this. Most reinforcement learning ideas were originally developed in the field of discounted Markov decision processes (Ross 1982). We describe these, and give a summary of recent results in reinforcement learning. Although similar results are available for the field of average reward Markov decision processes (MDPs), all the algorithms are simply modifications of those used for the discounted case, but with additional complications, and are not instructive for the purposes of this thesis. See Puterman (1994) for a description of these.

1.2.1 Discounted Markov decision processes

The study of Markov decision processes was initiated by Bellman (1952). This area involves strategic planning, with the need to balance out short and long term gains. A single agent makes an action choice at a series of discrete time steps, and moves through a series of states. The movement and the reward at each step are dependent on the action chosen. In the simplest (finite) case we have:

- a single agent,
- a finite set of states X,

- a finite set of actions $A(x)$ available at each state $x \in X$,
- a (possibly random) bounded reward $R(x,a)$ for each action at each state,
- given states x, y and an action a, a probability $P_{xy}(a)$ of being in state y at the next step.

A policy is a rule which the player uses to choose an action at each time step, and in general may depend on the time, the current state and the history of actions and states previously visited; an optimal policy maximises the expected discounted future reward. A stationary policy is a policy for which the current choice of action depends only on the current state: actions are selected according to a distribution $\pi(x) \in \Delta(x)$ when the agent is in state x (where $\Delta(x)$ denotes the set of probability distributions over the action set $A(x)$). A stationary policy is deterministic if $\pi(x)$ corresponds to a single action for each $x \in X$.

Theorem 5 (Bellman 1952) *The maximal expected discounted future reward starting from state x is given by the solution to the equations*

$$V(x) = \max_{a \in A(x)} \left\{ \mathbb{E}(R(x,a)) + \delta \sum_{y \in X} P_{xy}(a) V(y) \right\} \quad \textit{for each} \quad x \in X. \tag{1.11}$$

where $\delta \in (0,1)$ is the discount factor. This solution exists and is unique, and an optimal deterministic stationary policy exists and is given by the maximising action a in this equation.

Although the optimal policy is not necessarily unique (there may be more than one optimising action a), it should be noted that a deterministic stationary optimal policy does exist, and so we can restrict our attention to this class of policies.

It will be convenient, here and in the sequel, to define $r(x,a) = \mathbb{E}[R(x,a)]$ to be the expected immediate reward if action a is chosen in state x. Fixing a policy π, it is clear that the expected discounted future reward starting from state x and using policy π is given by the solution of the equations

$$V^{\pi}(x) = r(x, \pi(x)) + \delta \sum_{y \in X} P_{xy}(\pi(x)) V^{\pi}(y) \quad \text{for each} \quad x \in X, \tag{1.12}$$

where $r(x, \pi(x))$ and $P_{xy}(\pi(x))$ are defined in the obvious manner. This is just a set of $|X|$ linear equations in $|X|$ unknowns. A naive solution method would therefore be to find the solution to this equation for every possible deterministic policy π and choose the policy giving maximal expected reward. However this would be computationally intractable and more subtle methods need to be used.

1.2.2 Dynamic programming

The traditional approaches to solving discounted MDPs are collectively termed dynamic programming. Ross (1982) and Puterman (1994) provide introductions to these areas. There are essentially two different types of algorithm—value calculation algorithms, and policy improvement algorithms.

Value iteration

Define an operator $T : \mathbb{R}^{|X|} \to \mathbb{R}^{|X|}$ by

$$(TV)(x) = \max_{a \in A(x)} \left\{ r(x, a) + \delta \sum_{y \in X} P_{xy}(a) V(y) \right\} \quad \text{for each} \quad x \in X.$$

It is a simple calculation to show that T is a contraction mapping with respect to the supremum norm (i.e. $\|(TV) - (TV')\|_\infty \leq \|V - V'\|_\infty$ for two value functions V, V', with equality if and only if $V = V'$). The following theorem follows directly from this.

Theorem 6 *The iterates $T^n(V_0)$ converge uniformly to the unique solution to Bellman's equations (1.11), for any bounded initial conditions $V_0 \in \mathbb{R}^{|X|}$.*

This therefore provides a method of finding the optimal policy: find the solution to Bellman's equations (1.11) by finding the limit of the iterates of an arbitrary bounded V_0 under T, then choose a policy by using a maximising action at each state. This is known as value iteration, and generally requires an infinite number of iterations to find the optimal value function exactly. Stopping rules determining

when the estimated value function is close enough to use to choose an optimal policy are discussed by Puterman (1994).

Policy iteration

A different method is policy iteration, which follows the following algorithm.

1. Choose an initial policy π_0.

2. Evaluate V^{π_n} from equation (1.12).

3. For each $x \in X$, choose deterministic

$$\pi_{n+1}(x) \in \underset{\pi \in \Delta(x)}{\operatorname{argmax}} \left\{ r(x,\pi) + \delta \sum_{y \in X} P_{xy}(\pi) V^{\pi_n}(y) \right\},$$

   changing from $\pi_n(x)$ only if necessary.

4. If $\pi_{n+1} = \pi_n$ then this is an optimal policy, otherwise increment n and go to step 2.

Theorem 7 *For the policy iteration algorithm, $V^{\pi_{n+1}}(x) \geq V^{\pi_n}(x)$ for all $x \in X$, with equality at all $x \in X$ if and only if π_n is optimal. Thus the algorithm will converge in a finite number of steps (since X is finite and $A(x)$ is finite for each $x \in X$).*

Asynchronous dynamic programming

There is however a major computational issue for both value iteration and policy iteration if there are many states in the MDP (i.e. $|X|$ is large). An entire sweep of the whole state space is executed at each step, and the value function or policy is updated at every single state. This is true even if the inputs to that state are not modified (for example in an MDP where reward is only given out at a single state and there is strong structure in the transition matrices $P(a)$, the value will change at only a small number of stages on each iteration). This point has been addressed by asynchronous (or incremental) dynamic programming.

In this it is assumed that instead of updating the entire vector V at each stage of value iteration, or the entire policy π at each stage of policy iteration, only one component at a time will be updated. Thus for each iterate of value iteration we get

$$V_{n+1}(x_n) = (TV_n)(x_n) \text{ for some } x_n \in X$$

$$V_{n+1}(x) = V_n(x) \text{ for } x \neq x_n$$

and for each iterate of policy iteration we choose deterministic

$$\pi_{n+1}(x_n) \in \underset{\pi \in \Delta(x_n)}{\operatorname{argmax}} \left\{ r(x_n, \pi) + \delta \sum_{y \in X} P_{x_n y}(\pi) V^{\pi_n}(y) \right\}$$

for a single state $x_n \in X$ for which $\pi_n(x)$ is not maximising.

Theorem 8 (Williams and Baird 1993) *Asynchronous value iteration converges to the unique solution of Bellman's equations (1.11) so long as the value at each state is updated infinitely often. Asynchronous policy iteration will terminate in finite time at an optimal policy.*

Furthermore, the repeated solution of the linear equations (1.12) during policy iteration is costly in terms of computational time, and so hybrid schemes are considered where the value functions are updated using asynchronous value updates, and the policies are improved using asynchronous policy improvements. Here, at each stage, the pair (V_n, π_n) is transformed to (V_{n+1}, π_{n+1}) by either updating the value of a single state x_n:

$$V_{n+1}(x_n) = r(x_n, \pi_n(x_n)) + \delta \sum_{y \in X} P_{x_n y}(\pi_n(x_n)) V_n(y)$$

or by updating the policy at a single state:

$$\pi_{n+1}(x_n) \in \underset{\pi \in \Delta(x_n)}{\operatorname{argmax}} \left\{ r(x_n, \pi) + \delta \sum_{y \in X} P_{x_n y}(\pi) V_n(y) \right\}$$

for some x_n at which $\pi_n(x)$ does not already maximise this quantity. Again, convergence is shown:

Theorem 9 (Williams and Baird 1993) *Provided that*

$$r(x, \pi_0(x)) + \delta \sum_{y \in X} P_{\pi_0(x)}(x, y) V_0(y) \geq V_0(x)$$

for each $x \in X$, the (V_n, π_n) will converge to the optimal value function and policy if asynchronous value function updates and policy improvements are executed infinitely often for all states.

It should be noted that the initial condition is satisfied if the starting value function $V_0 = V^{\pi_0}$, so by picking an arbitrary initial policy π_0 and solving the linear equations (1.12) once to determine $V_0 = V^{\pi_0}$ we can guarantee convergence of this asynchronous hybrid scheme.

It clearly becomes an interesting problem to determine the best order in which to apply these asynchronous operators. Previous work in this area is usefully summarised in Chapter 9 of Sutton and Barto (1998). The two major ideas are those of prioritised sweeping, for which the policy or value is updated first at states x where there is most chance of there being a significant change (measured by the amount of change at states affecting the value of x since the last time an update was made at x). The other is to simulate the model, and to update states reached while playing the current policy. This will mean that actions and values affecting the current policy will be updated most frequently. Of course, with these approaches one needs to be careful to ensure that the all states are considered infinitely often.

1.2.3 Reinforcement learning algorithms

Reinforcement learning is a method for solving MDPs which is very closely related to asynchronous dynamic programming; Sutton and Barto (1998) and Bertsekas and Tsitsiklis (1996) provide a useful introduction to the area. However the concept was originally proposed as a model of animal learning (Thorndike 1898). The basic idea is that an agent experiments with actions and then uses the reward

received (and other information available) as reinforcement, either positive or negative, for the action played. A positively reinforced action will be more likely to be played again, whereas a negatively reinforced action will be less likely to be chosen.

Q-learning

Due to the close relationship with dynamic programming, the discounted reward case is again simpler and we restrict attention to this case. Consider Bellman's equations for discounted dynamic programming (1.11), and define

$$Q(x,a) = r(x,a) + \delta \sum_{y \in X} P_{xy}(a) V(y) \quad \text{for each} \quad x \in X, a \in A(x),$$

so that

$$V(x) = \max_{a \in A(x)} Q(x,a).$$

We can then rewrite (1.11) in terms of the Q values as

$$Q(x,a) = r(x,a) + \delta \sum_{y \in X} P_{xy}(a) \max_{b \in A(y)} Q(y,b) \quad \text{for each} \quad x \in X, a \in A(x). \quad (1.13)$$

It is clear that these equations are exactly equivalent to (1.11), and so everything that has been said about dynamic programming can be applied to these equations—in particular asynchronous backups can be performed, and the Q value at a single state-action pair can be updated at each step.

Suppose however that the transition matrices $P(a)$ and expected immediate rewards r are not known. By simply experimenting with different actions the agent can sample from the transition and reward distributions associated with those actions. Consider now the following algorithm, proposed by Watkins (1989), where at step n the agent is in state x_n, plays action a_n, receives reward R_n, and transitions to state x_{n+1}.

$$Q_{n+1}(x,a) = Q_n(x,a) + \lambda_{n+1} \mathbb{I}_{\{(x_n,a_n)=(x,a)\}} \left\{ R_n + \delta \max_{b \in A(x_{n+1})} Q_n(x_{n+1},b) - Q_n(x,a) \right\}$$

$$\text{for each} \quad x \in X, a \in A(x), \quad (1.14)$$

where $\lambda_{n+1} \in (0,1)$ is the learning parameter at step n. It is assumed throughout this section that the sequence $\{\lambda_n\}_{n\geq 1}$ of learning parameters satisfies

$$\lambda_n \in (0,1), \qquad \sum_{n\geq 1} \lambda_n = \infty, \qquad \sum_{n\geq 1} {\lambda_n}^2 < \infty. \tag{1.15}$$

This condition is standard throughout reinforcement learning, arising from the stochastic approximation theory used to prove the convergence theorems. Essentially it means that the process can move as far as necessary ($\sum_{n\geq 1} \lambda_n = \infty$) yet the variance will be bounded ($\sum_{n\geq 1} {\lambda_n}^2 < \infty$).

This can be seen as a stochastically sampled version of asynchronous value iteration.

Theorem 10 (Watkins and Dayan 1992) *The Q-learning algorithm (1.14) converges almost surely to the unique solution of the equations (1.13), provided that the Q value at each state-action pair is updated infinitely often.*

The theorem says that an agent in the environment may play any policy at all, and so long as all actions are played infinitely often at all states the learned Q values will solve the Bellman equations; an optimal policy will clearly be given by choosing a to maximise $Q(x,a)$ at each state x.

On-line learning

This raises the prospect of learning in real-time, in which an agent learns how to act optimally while continuing to move around and experiment in the environment. Clearly there is now a trade-off to be made between acting optimally according to current estimates of values, and experimenting to find potentially more rewarding actions—this is known as the exploration–exploitation trade-off. The two main ways of doing this are an ϵ-greedy scheme, and the smoother softmax scheme. For ϵ-greedy action choice, the action with greatest current value $Q_n(x_n,a)$ at state x_n will be played with probability $1-\epsilon$, but with probability ϵ a random action is chosen uniformly from $A(x_n)$. The softmax scheme is directly analogous to smooth

best responses (see Sections 1.1.2 and 3.2); the most common scheme is Boltzmann action selection, in which action a is played with probability

$$\frac{e^{Q_n(x_n,a)/\tau}}{\sum_{b\in A(x_n)} e^{Q_n(x_n,b)/\tau}}, \tag{1.16}$$

where τ is some positive temperature parameter—a low temperature will result in a distribution where the maximising action is played with probability close to 1, whereas a high temperature will result in a near uniform distribution over actions.

With online learning, the notion of asymptotic optimality is therefore important: a learning scheme is asymptotically optimal if, in the limit as $n \to \infty$, only optimal actions are played. If an exploration scheme is asymptotically optimal with respect to its Q value estimates (i.e. it plays the action a that maximises $Q_n(x_n, a)$), yet also guarantees infinitely many updates at each state action pair, then it is called "greedy in the limit with infinite exploration (GLIE)" (Singh *et al.* 2000). For this we need the probability of experimenting to decrease to zero in the limit, but to decrease slowly enough that all state-action pairs are visited infinitely often. Define a communicating MDP to be an MDP such that for each pair of states x, y there is a stationary policy π such that the probability of getting to y under π, starting at x, is greater than 0, and let $N_n(x)$ be the number of visits to state x up till step n of the learning process.

Theorem 11 (Singh *et al.* 2000) *The following exploration schemes are GLIE in a communicating MDP:*

1. *ϵ-greedy learning with the probability of playing an exploratory action given by $\epsilon_n = c/N_n(x_n)$ for some $c \in (0, 1)$.*

2. *Softmax learning with Boltzmann action choice where the temperature at step n is given by $\tau_n = (\max_a Q_n(x_n, a) - \min_a Q_n(x_n, a)) / \log N_n(x_n)$.*

Indeed if we assume the rewards, and hence all Q_n, are bounded then softmax exploration with Boltzmann action choice and temperatures given by $\tau_n =$

$C/\log N_n(x_n)$ for C a positive constant is GLIE in a communicating MDP. By following a GLIE scheme in a communicating MDP it is clear that the Q-learning algorithm will almost surely converge to the solution of (1.13) and the agent will play asymptotically optimally.

SARSA

Q-learning can be considered an off-policy algorithm, in that the updates made to Q values at each step depend not on the current policy, but on hypothetical actions which are optimal with respect to the current estimates Q_n (the update is made using the maximal Q value at the next state). This is the reason why the actual scheme used to choose actions is not important for the asymptotical convergence properties. On the other hand there is a variation of Q-learning under which the updates to the Q values are made according to the actual action chosen at the next state using the current policy. This is known as an on-policy algorithm, and the actual scheme used to choose actions is therefore crucial to the convergence properties of the algorithm.

The definition of the algorithm, known as SARSA since it updates according to the quintuple of State, Action, Reward, State, Action, is given by

$$Q_{n+1}(x,a) = Q_n(x,a) + \lambda_{n+1}\mathbb{I}_{\{(x_n,a_n)=(x,a)\}}\{R_n + \delta Q_n(x_{n+1},a_{n+1}) - Q_n(x,a)\}$$
$$\text{for each} \quad x \in X, a \in A(x), \tag{1.17}$$

This is identical to the Q-learning algorithm, except for the fact that the Q value from state x_{n+1} used to update $Q_n(x_n,a_n)$ is the Q value corresponding to the action chosen in the next state, as opposed to the maximal Q value for that state.

Theorem 12 (Singh *et al.* 2000) *Suppose a GLIE exploration scheme is used. Then almost surely the SARSA algorithm (1.17) will converge to the solution of the equations (1.13) and the agent will play asymptotically optimally.*

Although the convergence of Q-learning was originally proved using an esoteric argument (Watkins and Dayan 1992), the convergence of both these algorithms

has now been shown to arise directly from the traditional style of stochastic approximation result given in section 1.3.1. From that section, it will be clear that it is important that the discount factor δ is less than 1, which is what will give the stochastic contraction. A different way to get the stochastic contraction, without requiring a discount factor, is by assuming the task is episodic, i.e. will reach a terminal zero-reward state in finite time and then is restarted. This gives a stochastic contraction with respect to a weighted maximum norm. However we don't review these episodic tasks here.

$TD(\Lambda)$

The algorithms presented thus far "bootstrap", in that they use an arbitrary initial value function and make estimates based upon this arbitrary function (this is a feature of dynamic programming generally). However only one step of backup is performed at each stage (i.e. in SARSA only the Q value arising from the next state-action pair is used in an update). The natural question to ask is whether we could use the Q values at many subsequent states to provide information to the state-action pair currently being updated. This is achieved using the algorithm $TD(\Lambda)$ (Sutton 1988). This algorithm is merely a value estimating algorithm, calculating the state values for a fixed policy π.

Since advance knowledge of the future behaviour of the algorithm is not available, we use the following rule to feed back modifications to the value of the current state to update the value of previously visited states that have (possibly indirectly) used the value of the current state to bootstrap.

$$V_{n+1}(y) = V_n(y) + \lambda_{n+1}(R_n + \delta V_n(x_{n+1}) - V_n(x_n)) \sum_{m=0}^{n} (\delta\Lambda)^{n-m} \mathbb{I}_{\{\{x_m=y\}\}} \quad (1.18)$$

So the temporal difference $\lambda_{n+1}(R_n + \delta V_n(x_{n+1}) - V_n(x_n))$ which would be added to the value of $V_n(x_n)$ under an asynchronous stochastically sampled value iteration algorithm is instead fed back to all previous states visited, using a weighting parameter $\Lambda \in [0, 1]$ to determine how far back the difference is passed.

Theorem 13 (Jaakkola *et al.* 1994) *The $TD(\Lambda)$ algorithm (1.18) for $\Lambda \in [0,1]$ will converge almost surely to the correct value for the current policy provided that $\delta < 1$ and $\frac{(\lambda_{n+1}|x_n=x)}{\max_y(\lambda_{n+1}|x_n=y)} \to 1$ almost surely as $n \to \infty$, where $(\lambda_{n+1}|x_n = x)$ is the value of λ_{n+1} which would be used if the state at step n is $x_n = x$.*

In practise, instead of maintaining a history of the states visited in the past, eligibility traces are used. Define an eligibility trace recursively by

$$e_0(x) = 0 \quad \text{for all } x \in X$$

$$e_n(x) = \delta\Lambda e_{n-1}(x) + \mathbb{I}_{\{x_n=x\}}.$$

Then in vector notation we set

$$V_{n+1} = V_n + \lambda_{n+1}(R_n + \delta V_n(x_{n+1}) - V_n(x_n))e_n.$$

This is identical to the algorithm (1.18) but provides an online method where only the eligibility vector e_n and the value vector V_n need to be stored. Singh and Sutton (1996) modify this approach by instead setting $e_n(x_n) = 1$, and claim to improve the robustness of the algorithm. They call this "replacing eligibility traces" (as opposed to the "accumulating eligibility traces" described above). The algorithm will converge under the same conditions as for accumulating illegibility traces.

Note however that this is merely a value calculation method which will return V^π for a particular π. It is therefore not a solution method for MDPs, and should be regarded as a temporal difference method to calculate the value of a policy. However it has proved useful to Tesauro (1994), who has created a backgammon player that learns to play at master-level based on the $TD(\Lambda)$ algorithm.

Several true control algorithms have been suggested using eligibility traces, including SARSA(Λ) (Rummery 1995) and two proposed extensions of off-policy Q-learning to this domain (Sutton and Barto 1998, Chapter 7). None of these algorithms have been proven to converge.

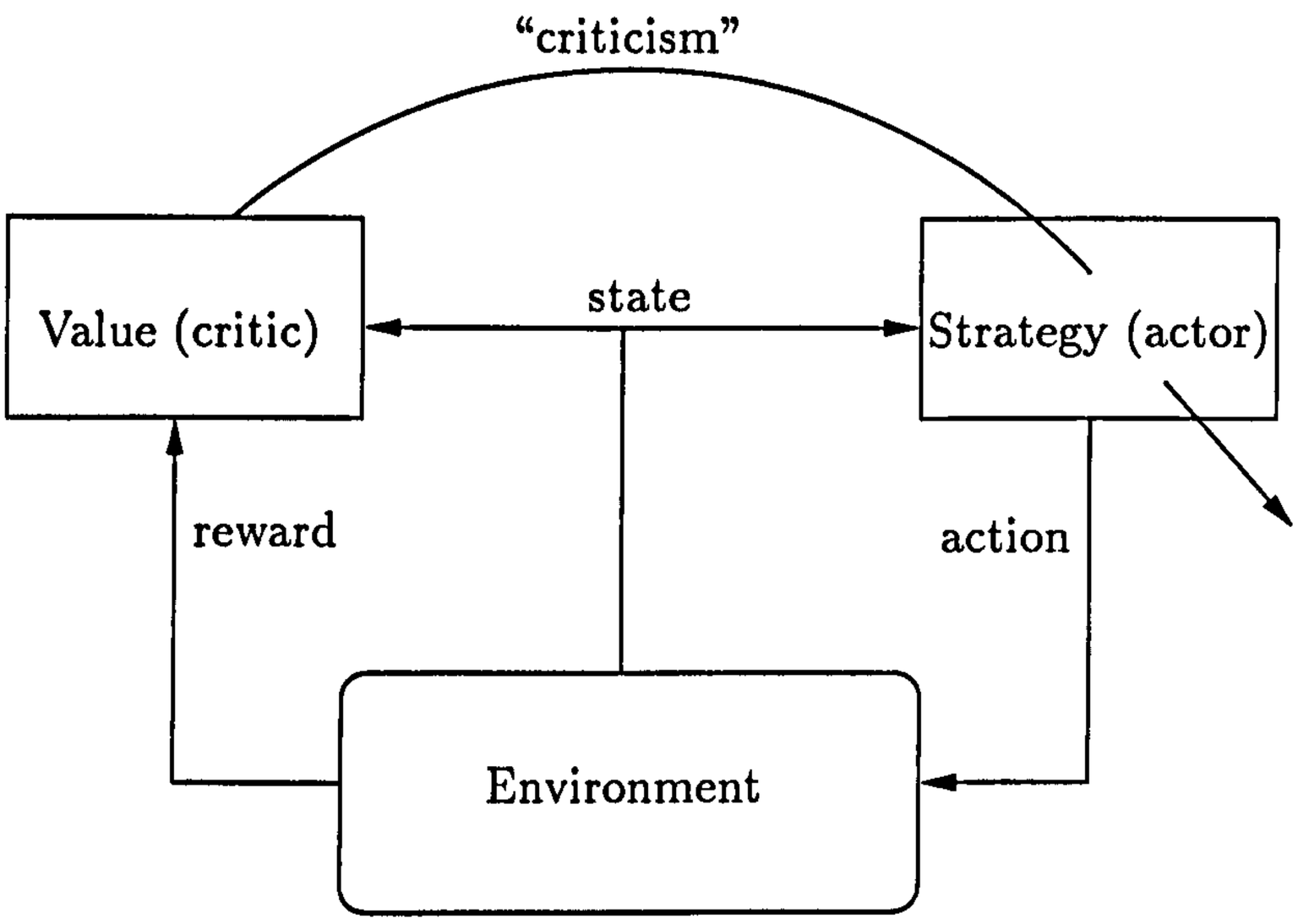


Figure 1.1: Schematic diagram of actor–critic algorithms.

Actor–critic algorithms

Another class of reinforcement learning algorithms is actor–critic methods. In these the algorithm maintains a separate policy (the actor) and value function (the critic). The critic then provides information about the value of actions, and this is used by the actor to improve the policy. A schematic diagram is given in Fig. 1.1. This is related to the policy iteration algorithm, where the value function V^π acts as a critic to the actor π. Although closely related to the hybrid asynchronous value/policy iteration schemes described above, the general properties of these algorithms are not known. However Crites and Barto (1995) have described an actor-critic algorithm which is equivalent to Q-learning, and Konda and Borkar (2000) have proved the convergence an actor–critic algorithm which is an approximated version of policy iteration (in the same manner as Q-learning is an approximation of value iteration).

Function approximation

An aspect of reinforcement learning which has not been examined here, but is covered by Bertsekas and Tsitsiklis (1996), is that of function approximation. In many problems the size of the state space is huge, and too big for a table of all V or Q values to be stored, so a functional representation is used as an approximation.

1.3 Stochastic approximation

The subject of stochastic approximation was introduced as an iterative method to find the zeroes of a function when only a perturbed value of the function at the current estimate of the zero is known (Robbins and Monro 1951). However it has found application in various modern disciplines: in neural networks it is used to prove convergence of stochastic backpropagation methods, while in reinforcement learning it is the tool used to provide most of the convergence results in the literature. Stochastic approximation will be used to prove many of the results of this thesis.

In general we examine an algorithm of the form

$$\theta_{n+1} = \theta_n + \lambda_{n+1} F_{n+1}(\theta_n), \tag{1.19}$$

where $\{F_n\}_{n\geq 1}$ is a sequence of random functions and $\{\lambda_n\}_{n\geq 1}$ is a sequence of learning parameters. Throughout this section, we will assume that $\{\mathcal{F}_n\}_{n\geq 0}$ is an increasing sequence of σ-fields such that $\mathcal{F}_n$ contains the stochastic approximation process up till time n.

There are two approaches in the current literature. The first is the Robbins-Monro style, in which conditions placed on $\{F_n\}_{n\geq 1}$ and $\{\lambda_n\}_{n\geq 1}$ ensure directly that $\theta_n \to 0$ almost surely. The second approach, known as the ODE approach (standing for ordinary differential equation), was suggested by Ljung (1977) and developed into a fruitful theory by Kushner and Clark (1978). For this approach the iterates of (1.19) are shown to approximate the solution of a related differen-

tial equation, and convergence properties of the stochastic approximation can be inferred from the behaviour of solutions of the differential equation.

1.3.1 Robbins–Monro style algorithms

There have been several generalisations of the original algorithm, many of which have been made by researchers in reinforcement learning. All of the results using this approach assume that the F_n satisfy contraction properties of some sort. In reinforcement learning the usual contraction property is with respect to a weighted maximum norm. The weighted maximum norm of a vector $x \in \mathbb{R}^d$ is defined by

$$\|x\|_W = \max_{i=1,\dots,d} \frac{|x(i)|}{W(i)},$$

where the weight vector W has positive components. The reason this norm is introduced is because for episodic tasks the value iteration function is a contraction for a weighted maximum norm even when the discount factor $\delta = 1$.

The following theorem provides a recent result using this method, in which the learning parameters can depend on the component of θ_n. This allows for the study of asynchronous algorithms, where only some components of θ are updated for each n.

Theorem 14 (Singh *et al.* 2000) *A random iterative process*

$$\theta_{n+1}(z) = (1 - \lambda_{n+1}(z))\theta_n(z) + \lambda_{n+1}(z)G_{n+1}(\theta_n)(z)$$

converges to zero with probability 1 if the following properties hold:

1. *The set of possible states z is finite,*
2. $0 \leq \lambda_n(z) \leq 1$, $\sum_{n\geq 1} \lambda_n(z) = \infty$, $\sum_{n\geq 1} {\lambda_n}^2(z) < \infty$ *a.s.,*
3. $\|\mathbb{E}(G_{n+1}(\theta_n)|\mathcal{F}_n)\|_W \leq \kappa\|\theta_n\|_W + c_n$ *where* $\kappa \in [0,1)$ *and* $c_n \to 0$ *a.s.,*
4. $Var(G_{n+1}(\theta_n)(z)|\mathcal{F}_n) \leq K(1 + \|\theta_n\|_W)^2$, *where K is a constant.*

Here W is a weight vector and $\mathcal{F}_n$ is an increasing sequence of σ-fields including the past of the process; in particular we assume that $\lambda_n, \theta_n, G_n \in \mathcal{F}_n$.

The asynchronicity is apparent in the Q-learning process, for which $\lambda_{n+1}(x,a) = \lambda_{n+1}\mathbb{I}_{\{(x_n,a_n)=(x,a)\}}$ is non-zero only if $(x,a) = (x_n,a_n)$. In this case the condition $\sum_n \lambda_n(z) = \infty$ is effectively a condition saying that the updates must be performed infinitely often at all states.

Singh *et al.* (2000) use this theorem to prove the convergence of SARSA, whereas Jaakkola *et al.* (1994) use a slight variation to prove the convergence of $TD(\Lambda)$. This is also essentially the same theorem used by Littman and Szepesvári (1996) to prove that joint action maximin Q-learning will converge in 2-player zero-sum stochastic games (see Section 1.4)—various other generalised Q-learning algorithms for discounted reward MDPs are also proved to converge in the same paper.

1.3.2 The ODE Approach

Although the result of the previous section proves useful in certain instances of reinforcement learning, there are many situations under which such a contraction property does not hold. For these, a more general method must be employed. The ODE (ordinary differential equation) method of stochastic approximation is based on the observation that if $F_n(\theta) = F(\theta) + U_n$, where F is a Lipschitz continuous function and $\{U_n\}_{n\geq 1}$ is a sequence of random variables with bounded variation and mean 0, then (1.19) is a noisy discretisation of the ODE

$$\dot{\theta} = F(\theta). \tag{1.20}$$

The limiting behaviour of the stochastic approximation process is therefore related to the asymptotic behaviour of trajectories of (1.20). This approach was first explored by Ljung (1977), but is generally attributed to Kushner and Clark (1978). Recent developments are in Kushner and Yin (1997), with a more readily applicable set of results in Benaïm (1999). Benaïm and Hirsch (1999) have used this

approach to study the behaviour of stochastic fictitious play, relating it to the smooth best response dynamics (1.8). We will extend their approach in several directions in this thesis. We follow the approach of Benaïm (1999), starting with some standard definitions about trajectories of dynamical systems; we assume that a dynamical system is defined on a metric space (M, d), with induced norm $\|\cdot\|$.

Definition 15

1. *A semiflow φ on a metric space (M, d) is a continuous map $\varphi : \mathbb{R}_+ \times M \to M$, $(t, x) \mapsto \varphi_t(x)$, such that $\varphi_0 = Identity$ and $\varphi_{t+s} = \varphi_t \circ \varphi_s$. An ODE on M induces a semiflow φ.*

2. *An invariant set for the semiflow φ is a set $S \subset M$ such that $\varphi_t(S) = S$ for all $t \geq 0$.*

3. *$S \subset M$ is an attractor for φ if S is non-empty, compact, invariant, and has a neighbourhood $U \subset M$ such that $d(\varphi_t(x), S) \to 0$ as $t \to \infty$ uniformly in $x \in U$.*

4. *The basin of attraction $B(S)$ of an attractor S is the set of points x such that $d(\varphi_t(x), S) \to 0$ as $t \to \infty$. S is globally attracting if $B(S) = M$.*

5. *Let S be a compact invariant set of the semiflow φ. A continuous function $V : M \to \mathbb{R}$ is a Lyapunov function for S under φ if $V(\varphi_t(x))$ is constant for $x \in S$ and strictly decreasing in t whenever $x \notin S$.*

6. *Given a trajectory $X : \mathbb{R}_+ \to M$, the (omega) limit set of the trajectory is the set $\{x \in M : X(t_k) \to x$ for some sequence $t_k \to \infty\}$.*

The central concept of Benaïm's approach to stochastic approximation is the asymptotic pseudotrajectory. This is a trajectory in M which asymptotically tracks the solutions of (1.20).

Definition 16 *A continuous function $X : \mathbb{R}_+ \to M$ is an asymptotic pseudotrajectory for a semiflow φ if, for any $T > 0$,*

$$\lim_{t\to\infty} \sup_{0\le h\le T} d(X(t+h), \varphi_h(X(t))) = 0.$$

Proposition 17 (Benaïm 1999) *Consider a general stochastic approximation process (1.19) in the metric space (M, d). Let $F_{n+1}(\theta_n) = F(\theta_n) + U_{n+1}$, where F is a globally integrable vector field, the U_n are perturbations, and $\{\lambda_n\}_{n\ge1}$ satisfies*

$$\sum_{n\ge1} \lambda_n = \infty, \quad \lambda_n \to 0 \quad \text{as } n \to \infty.$$

Let $\tau_n = \sum_{k=1}^{n} \lambda_k$ and $m(t) = \sup\{n \ge 0 : t \ge \tau_n\}$, and assume that

1. *For all $T > 0$,*

$$\lim_{n\to\infty} \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1} U_{l+1} \right\| : k = n+1, \ldots, m(\tau_n + T) \right\} = 0.$$

2. $\sup_n \|\theta_n\| < \infty$.

Then an interpolation of the θ_n is an asymptotic pseudotrajectory of the semiflow induced by (1.20).

Define the limit set of a stochastic approximation process $\{\theta_n\}_{n\ge0}$ to be the set

$$L(\{\theta_n\}) = \{\theta \in M : \theta_{n_k} \to \theta \quad \text{for some subsequence } n_k\}$$

Since $L(\{\theta_n\})$ is contained in the limit set of any interpolation, it follows that the limit set of the (random) stochastic approximation process is contained the limit set of an asymptotic pseudotrajectory of the (deterministic) ODE (1.20). It therefore suffices to consider the conditions of Proposition 17, and then to characterise the limit set of asymptotic pseudotrajectories. We present a useful proposition giving conditions under which Condition 1 of Proposition 17 holds.

Proposition 18 (Benaïm 1999) *Suppose that $\{\lambda_n\}_{n\geq 1}$ is a deterministic sequence, that $U_{n+1} = M_{n+1} + b_{n+1}$ where $\{M_n\}_{n\geq 1}$ is adapted with respect to $\mathcal{F}_n$ and $b_n \to 0$ as $n \to \infty$, and that $\mathbb{E}(M_{n+1} \mid \mathcal{F}_n) = 0$. Suppose also that for some $q \geq 2$*

$$\sup_{n\geq 1} \mathbb{E}(\|M_{n+1}\|^q) < \infty, \quad \textit{and} \quad \sum_{n\geq 1} {\lambda_n}^{1+q/2} < \infty.$$

Then assumption 1 of Proposition 17 holds with probability 1.

Note that if $q = 2$ in this theorem, then the conditions on the learning parameters are identical to the conditions (1.15) stated in the section on reinforcement learning, and the resulting conditions on the M_n are for them to have bounded second moments.

Assumption 2 of Proposition 17 (on boundedness of the iterates) causes problems for many applications of stochastic approximation theory. Solutions are to be found in the projections of Kushner and Clark (1978), or in the random truncations of Chen and Zhu (1986). However, in most of our applications the iterates will be bounded automatically by the nature of the problem.

In order to characterise the limit set of asymptotic pseudotrajectories, we introduce the following:

Definition 19 *Given a semiflow φ, define an (ϵ, T)-pseudotrajectory from $x \in M$ to $y \in M$ to be a sequence of points $(y_0, \ldots, y_k)$ and times $(t_1, \ldots, t_k)$ with each $t_j \geq T$ such that*

$$d(y_0, x) < \epsilon, \quad d(\varphi_{t_j}(y_j), y_{j+1}) < \epsilon, \quad \textit{and} \quad y_k = y.$$

Call x a chain-recurrent point of φ if there is an (ϵ, T)-pseudotrajectory from x to x for all positive T and ϵ. A compact invariant set $L \subset M$ of the semiflow φ is called internally chain-recurrent if every point $p \in L$ is chain-recurrent for $\varphi|L$ (the semiflow restricted to the set L).

Note the similarity between this concept of chain-recurrence and the cyclically stable sets of Gilboa and Matsui (1991).

Theorem 20 (Benaïm 1999) *Let $X : \mathbb{R}^+ \to M$ be an asymptotic pseudotrajectory of the semiflow φ. Then the limit set of X is a connected, compact, internally chain-recurrent, invariant set of the semiflow φ.*

Combining Proposition 17 with Theorem 20 shows that the limit set of a stochastic approximation process is an internally chain recurrent set for the semiflow induced by the ODE (1.20).

Corollary 5.4 and Proposition 6.4 of Benaïm (1999) provide easily verifiable conditions which can be used to characterise the chain-recurrent sets of a flow, and hence the limit sets of the asymptotic pseudotrajectories:

Proposition 21 (Benaïm 1999) *Let L be the limit set of an asymptotic pseudotrajectory of the semiflow φ, and let Λ be a compact invariant set of φ.*

1. *If Λ is a global attractor of φ, then $L \subset \Lambda$.*
2. *Let V be a Lyapunov function for Λ under φ. If $V(\Lambda) \subset \mathbb{R}$ has empty interior then $L \subset \Lambda$ and $V(L)$ is constant.*

Benaïm (1999) shows that all attractors will contain the limit set with positive probability, provided that the stochastic approximation process can enter their basin of attraction with positive probability. In contrast to this result, we wish to rule out certain parts of the chain recurrent set; this set is a superset of the recurrent set, and therefore contains all equilibria and periodic orbits. Since the perturbations *allowed* in the definition of chain recurrence are *enforced* by the stochasticity of the approximation algorithm, it would seem reasonable to expect that linearly unstable equilibrium points and periodic orbits will have zero probability of containing the limit set of the approximation process. This is indeed the case under conditions placed on the noise at these points. The basic result is due to Pemantle (1990), but has been studied further by Benaïm (1999) and Brandière (1998a, 1998b).

Theorem 22 (Pemantle 1990) *Let θ be an equilibrium point of the semiflow φ, and let U be a neighbourhood of 0. Assume there are constants $\rho \in (\frac{1}{2}, 1]$ and $c_1, c_2, c_3, c_4 > 0$ such that the following conditions are met whenever $\theta_n \in U$ and n is sufficiently large:*

- *0 is a linearly unstable critical point of the semiflow,*
- $\frac{c_1}{n^\rho} \leq \lambda_n \leq \frac{c_2}{n^\rho}$,
- $\mathbb{E}((U_{n+1} \cdot v)^+ \,|\, \mathcal{F}_n) \geq c_3$ *for each unit vector v in the codomain of F,*
- $\|U_n\| \leq c_4$,

where $(U_{n+1} \cdot v)^+$ is the positive part of $U_{n+1} \cdot v$. Assume also that F is smooth enough to apply the stable manifold theorem (at least C^2). Then $\mathbb{P}(\theta_n \to \theta) = 0$.

The asynchronous case considered in Theorem 14, where $\lambda_n(z)$ is state-dependent, is more difficult to deal with using the ODE approach, since when different components are updated at different rates the differential equation (1.20) will need to be modified to take this into account. Kushner and Yin (1997) have analysed the algorithm using a weak convergence criteria and the theory of differential inclusions. Almost sure convergence is studied by Borkar (1998).

1.3.3 Two-timescales stochastic approximation

Borkar (1997) studies concurrent approximation processes with learning parameters that approach zero at different rates. Essentially what this means is that there is a 'fast' process, which can be considered to be always fully calibrated to the 'slow' process. However, both processes evolve concurrently, and it is only the magnitude of the learning parameters that differs. We here present a slight reformulation of the result of Borkar (1997) which places the result in the framework of Benaïm (1999). This is a special case of Theorem 40, and the proof is reserved till Chapter 4.

Theorem 23 (Borkar 1997) *Consider two coupled stochastic approximation processes*

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} + \lambda_{n+1}^{(i)} \left\{ F^{(i)}(\theta_n^{(1)}, \theta_n^{(2)}) + U_{n+1}^{(i)} \right\}, \quad \textit{for} \quad i = 1, 2,$$

where, for each i, the following conditions hold:

B1 $\theta^{(i)} \in M^{(i)}$, *where* $(M^{(i)}, d^{(i)})$ *is a compact metric space,*

B2 $F^i : M^{(1)} \times M^{(2)} \to M^{(i)}$ *is Lipschitz,*

B3 $\sum_{n \geq 1} \lambda_n^{(i)} = \infty$ *and* $\lambda_n^{(i)} \to 0$ *as* $n \to \infty$, *and*

B4 *for all* $T > 0$,

$$\lim_{n \to \infty} \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(i)} U_{l+1}^{(i)} \right\| : k = n+1, \ldots, m^i(\tau_n^{(i)} + T) \right\} = 0,$$

where we define $\tau_n^{(i)} = \sum_{k=1}^{n} \lambda_k^{(i)}$ *and* $m^{(i)}(t) = \sup\{n \geq 0 : t \geq \tau_n^{(i)}\}$.

Further, the $\lambda_n^{(i)}$ *satisfy*

$$\lambda_n^{(1)} / \lambda_n^{(2)} \to 0 \quad \textit{as} \quad n \to \infty.$$

Suppose that for each $\theta^{(1)}$ *the ODE*

$$\dot{Y} = F^{(2)}(\theta^{(1)}, Y) \tag{1.21}$$

has a unique globally asymptotically stable equilibrium point $\xi(\theta^{(1)})$ *such that* ξ *is Lipschitz. Then, almost surely,*

$$\|\theta_n^{(2)} - \xi(\theta_n^{(1)})\| \to 0 \quad \textit{as} \quad n \to \infty$$

and a suitable continuous time interpolation of the process $\{\theta_n^{(1)}\}_{n \geq 0}$ *is an asymptotic pseudotrajectory of the semiflow induced by the ODE*

$$\dot{X} = F^{(1)}(X, \xi(X)). \tag{1.22}$$

As observed by Borkar (1997), the condition **B4** is satisfied whenever the $\{\lambda_n^{(i)}\}_{n \geq 1}$ are deterministic sequences with $\sum_{i \geq 1} (\lambda_n^{(i)})^2 < \infty$ and the $M_n^{(i)}$ are martingale differences with bounded second moment (this is closely related to Proposition 18).

1.4 Stochastic games

Stochastic (or Markov) games were initially introduced by Shapley (1953), in the context of zero sum games; a comprehensive survey is given by Filar and Vrieze (1997). They are a generalisation of both normal form games and Markov decision processes; players must act strategically, balancing out their immediate and long term rewards, but these rewards also depend on the actions of the other players. The technical framework is:

- Λ finite set of players $\{1, \ldots, N\}$,
- Λ finite set of states X,
- Λ finite set of actions $A^i(x)$ available to each player $i = 1 \ \ldots, n$ at each state $x \in X$, resulting in a finite set of joint actions $\underline{A}(x) = A^1(x) \times \ldots \times A^N(x)$ at each state.
- Λ (possibly random) bounded reward $R^i(x, \underline{a})$ awarded to player i when joint action $\underline{a} \in \underline{A}(x)$ is played in state $x \in X$; as for MDPs, define $r^i(x, \underline{a}) = \mathbb{E}[R^i(x, \underline{a})]$,
- Given states x, y and joint action $\underline{a}$, a probability $P_{\underline{a}}(x, y)$ of being in state y at the next step.

We see that this is a direct combination of both normal form games and Markov decision processes, including both as a special case (reducing the player set to a singleton gives an MDP, while reducing the state set to a singleton plus an absorbing zero-reward state gives normal form games). Writing $\Delta^i(x)$ for the set of mixed strategies of player i at state x, and letting

$$V^i(x) = \max_{\pi^i \in \Delta^i(x)} \min_{a^{-i} \in A^{-i}(x)} \sum_{a^i \in A^i(x)} \pi^i(a^i) Q^i(x, (a^i, a^{-i}))$$

$$Q^i(x, \underline{a}) = r^i(x, \underline{a}) + \delta \sum_{y \in X} P_{\underline{a}}(x, y) V^i(y), \quad \text{for each } x \in X, \underline{a} \in \underline{A}(x).$$

Shapley (1953) shows that for 2-player zero-sum stochastic games, taking maximin strategies in place of the max used in dynamic programming (Section 1.2) again gives a contraction mapping if the rewards are discounted. Then the same arguments as used in dynamic programming imply that a stationary strategy will suffice, and that a solution exists. Filar and Vrieze (1997) generalise this result to show that for general games, if rewards are discounted, a Nash equilibrium must exist in stationary strategies. However calculating the equilibrium strategies may not be easy.

1.4.1 Reinforcement learning in stochastic games

Littman (1996) extends Shapley's idea to reinforcement learning, thus generalising Q-learning to the area of discounted 2-player zero-sum stochastic games. Let

$$V_n^1(x) = \max_{\pi^1 \in \Delta^1(x)} \min_{a^2 \in A^2(x)} \sum_{a^1 \in A^1(x)} \pi^1(a^1) Q_n^1(x, (a^1, a^2))$$

$$Q_{n+1}^1(x, \underline{a}) = Q_n^1(x, \underline{a}) + \lambda_{n+1} \mathbb{I}_{\{(x_n, \underline{a}_n) = (x, \underline{a})\}} \left\{ R_n^1 + \delta V_n^1(x_{n+1}) - Q_n^1(x, \underline{a}) \right\}$$

$$\text{for each } x \in X,\ \underline{a} \in \underline{A}(x),$$

with V_n^2 and Q_n^2 defined analogously. This algorithm (called minimax-Q) converges to the Nash equilibrium of discounted 2-player zero-sum stochastic games under similar conditions to those imposed on single-agent Q-learning (Littman 1996).

However, it requires that both players can observe their opponent's action choices, and also that they can solve a matrix game each time they play. The first requirement may or may not be satisfied, depending on the situation the players find themselves in; the second requires that players solve a matrix game (which is a linear programming problem) every time an action must be played. This is against the spirit of reinforcement learning, where agents should not make any complex calculations based upon knowledge of the environment.

Hu and Wellman (1998) have attempted to extend this result to general-reward 2-player games. However they place restrictive assumptions on the Q values ob-

tained at every step of the learning process—something which would be very difficult to guarantee in practise (Bowling 2000).

Claus and Boutilier (1998) formally observe the difference between "joint action learners", who observe the actions and rewards of all players, and "individual learners", where players only respond to the rewards they receive. Littman's minimax-Q is an example of joint action learners, while the algorithms of most interest in this thesis are individual learners.

Bowling and Veloso (2002) introduce WoLF learning, an individual learning algorithm for general games. Here, players record the average reward they have received, and modify their learning rates depending on whether they perceive their current strategy as performing better or worse than the historical average. They show that a version for 2-player normal form games where each player has only two actions will converge to Nash equilibrium. Further theoretical results are not available, but the empirical behaviour of the algorithm is impressive, even for complex stochastic games.

Borkar (2001) provides the most advanced theoretical analysis of individual learners in stochastic games, with a generalisation of the actor–critic algorithm of Konda and Borkar (2000) to the multi-agent setting. He shows that appropriate empirical mixed strategies arising from the process (though not the usual Cesaro sums of actions played) converge to a generalised Nash equilibrium, a concept developed in the paper and related to correlated Nash equilibrium.

1.5 Motivating remarks

As already observed, there is much interest in whether Nash equilibrium play can be explained by some process other than introspective analysis by the players. Many previous attempts have either used a population view of mixed strategies (in the evolutionary literature) or have assumed a knowledge of the reward structure of the game. In this thesis we will study reinforcement learning algorithms for games,

in which agents require no prior knowledge of the game; these are analogous to, and mainly inspired by, reinforcement learning algorithms for Markov decision processes.

Part of the motivation for the study of such processes is the observation that in many biological and economic applications of game theory the participants do not know that they are playing a game, let alone know the structure of the game, yet equilibrium play can still be observed. Examples of this approach are given by Hofbauer and Sigmund (1998) and Fudenberg and Levine (1998).

A separate inspiration is the need for multi-agent learning in complex control environments (Crites and Barto 1998; Boyan and Littman 1994; Singh and Bertsekas 1997), as well as agents that learn to play games (in the lay sense of the word) through self play (Tesauro 1994; Schraudolph *et al.* 1994; Stone 2000).

Many of these applications are actually in the field of stochastic games. However, the theory of reinforcement learning in normal form games is still far from complete—previous analytical work in this area is largely restricted to the consistent reinforcement learning algorithms described in Section 1.1.4, for which the empirical distributions of play will converge to a correlated Nash equilibrium, and some results on 2×2 games (Bowling and Veloso 2002; Singh *et al.* 2000). This thesis will concentrate largely on normal form games, since these provide a (relatively) simple setting where ideas can be developed and studied theoretically. It is anticipated that these ideas and theoretical results will be extended to stochastic games in the future.

Reinforcement learning is interesting because players must discover how to act in an unknown environment. On the other hand, the environment faced by any particular player of a game is not stationary, because opponents are also learning. By using sophisticated tools from the theory of stochastic approximation and dynamical systems we can analyse the asymptotic behaviour of these processes, determining whether (or not) convergence will occur in the long term.

1.6 Outline of the thesis

We start, in Chapter 2, by considering Börgers and Sarin's (1997) version of a 'mathematical model for simple learning' (Bush and Mosteller 1951). This results in a stochastic approximation of the replicator dynamics (1.9). We describe some of the problems inherent to learning in games, and introduce a new algorithm which uses two-timescales stochastic approximation to approximate the payoff-normalised replicator dynamics (1.10).

In Chapter 3 we introduce an actor–critic learning algorithm in which players update their estimates of the value of actions on a faster timescale than they adapt their strategies towards a smooth best response to these estimates. This means that the value estimates can be considered to be accurate, and the strategies are a stochastic approximation of the smooth best response dynamics (1.8). A related population process is also considered.

The actor–critic algorithm is extended in Chapter 4, and players now all adapt their strategies on different timescales. This necessitates an extension of Borkar's two-timescales stochastic approximation result (Theorem 23) to multiple timescales. The algorithm is shown to converge to Nash distributions for several classes of game, and in particular for two games which have caused difficulty for most (if not all) previous learning processes.

The fundamental algorithm of reinforcement learning in MDPs is Q-learning (Watkins 1989). We study a version of this for normal form games in Chapter 5. It is shown that the Q values evolve in a manner closely related to the smooth best response dynamics, and thus will converge in 2-player zero-sum games. A modification of this algorithm where the players learn at different rates is also considered.

Chapter 6 introduces a version of the actor–critic algorithm of Chapter 3 where players adapt towards a best response (instead of a smooth best response). This algorithm is studied by analysing a generalisation of weakened fictitious play pro-

cesses (Van der Genugten 2000), relating the behaviour to that of the best response dynamics (1.7). Our discontinuous actor–critic algorithm is the first reinforcement learning algorithm proven to converge to Nash equilibrium in general 2-player zero-sum games, and N-player partnership games, as well as other classes of games.

Attempts to extend these algorithms into stochastic games have demonstrated that it will be useful to consider carefully the nature of smooth best response functions. In Chapter 7 a method is developed for which there is a unique smooth best response to opponent strategies in a stochastic game.

Further work is suggested in Chapter 8.

Chapter 2

A model for simple learning

In this chapter we modify a reinforcement learning algorithm for repeated normal form games, known as "stimulus-response learning" (Börgers and Sarin 1997). This algorithm is a special case of Bush and Mosteller's (1951) "mathematical model for simple learning", and has also been studied in the context of automata games (Narendra and Thathachar 1989). Börgers and Sarin (1997) show that each player will converge to a pure strategy with probability 1. This is an obvious failure of the algorithm, since then for certain games (those with only mixed equilibria) convergence to the Nash equilibrium has zero probability.

In this chapter we argue that Börgers and Sarin's use of a fixed learning parameter is partly to blame for this behaviour, and that by introducing a variable learning parameter satisfying the standard conditions (1.15) an improvement is gained. We show that an interpolation of our algorithm is an asymptotic pseudo-trajectory of the replicator dynamics (1.9), and that if the process converges to a mixed strategy then it must be a Nash equilibrium of the game.

Section 2.1 presents our model and the basic stochastic approximation result, then Section 2.2 provides a brief summary of the known results about the replicator dynamics. In Section 2.3 we apply the algorithm to some simple games, while in Section 2.4 a modification of the algorithm which relates to the normalised replicator dynamics (1.10) is examined.

2.1 Stimulus–response learning

We model learning in a repeated normal form game using a very simple algorithm under which players observe a stimulus (the reward received) and produce a response (a direct modification of their strategy). Although this mixes rewards directly with policies, when in reality these quantities exist in different spaces, it results in a model which can be analysed easily. Suppose that at the nth play of the game, player i uses mixed strategy π_n^i to select action a_n^i and receives reward $R_n^i \in [0,1]$. This is clearly a restricted class of games, but rewards of any game with bounded rewards can be rescaled to fit in this framework.

After game n, player i updates their strategy according to

$$\pi_{n+1}^i(a^i) = (1 - \lambda_{n+1} R_n^i)\pi_n^i(a^i) + \lambda_{n+1} R_n^i \mathbb{I}_{\{a_n^i = a^i\}} \quad \text{for all} \quad a^i \in A^i, \qquad (2.1)$$

where $\{\lambda_n\}_{n\geq 1}$ is a deterministic sequence satisfying the standard reinforcement learning conditions (1.15), and also satisfying $\lambda_n \in (0,1)$. This is identical to the algorithm proposed by Börgers and Sarin (1997), except for the fact that the learning parameters λ_n vary with n.

Writing $\underline{a}_n$ for the joint action played at game n, π_n for the joint mixed strategy used at game n, and π_n^{-i} for the opponent mixed strategy at game n (see Section 1.1), and assuming that the R_n^i arise from payoffs in a game, we see that $\mathbb{E}[R_n^i \,|\, \underline{a}_n] = r^i(\underline{a}_n)$, $\mathbb{E}[R_n^i \,|\, a_n^i, \pi_n^{-i}] = r^i(a_n^i, \pi_n^{-i})$, and $\mathbb{E}[R_n^i \,|\, \pi_n] = r^i(\pi_n)$. Note that the R_n^i could be independent random variables with appropriate means without enforcing any change to our analysis. Thus we see that

$$\begin{aligned}
\mathbb{E}[\pi_{n+1}^i(a^i) - \pi_n^i(a^i) \,|\, \pi_n] &= \lambda_{n+1} \sum_{\underline{b} \in \underline{A}} \pi_n^1(b^1) \cdots \pi_n^N(b^N) r^i(\underline{b}) \left\{ \mathbf{I}_{\{b^i = a^i\}} - \pi_n^i(a^i) \right\} \\
&= \lambda_{n+1} \sum_{b^i \in A^i} \pi_n^i(b^i) r^i(b^i, \pi_n^{-i}) \left\{ \mathbf{I}_{\{b^i = a^i\}} - \pi^i(a^i) \right\} \\
&= \lambda_{n+1} \pi_n^i(a^i) \left\{ r^i(a^i, \pi_n^{-i}) - r^i(\pi_n) \right\}.
\end{aligned}$$

Defining $F(\pi)^i(a^i) = \pi^i(a^i) \left\{ r^i(a^i, \pi^{-i}) - r^i(\pi) \right\}$ it is clear that

$$\pi_{n+1}^i = \pi_n^i + \lambda_{n+1}^i \left(F(\pi_n) + U_{n+1} \right)$$

where U_{n+1} has zero mean and is bounded (since $R_n^i \in [0,1]$). Therefore Proposition 18 holds, the conditions of Proposition 17 are satisfied, and Theorem 20 shows the following:

Theorem 24 *With probability 1 the limit set L of the learning process defined in (2.1) is a connected, compact, internally chain-recurrent, invariant set for the semiflow φ defined by the asymmetric replicator dynamics (1.9).*

2.2 Replicator dynamics

Thus we wish to use the results of Section 1.3 to characterise the limit set of our learning algorithm, by determining the internally chain-recurrent sets for the semiflow induced by the replicator dynamics. These dynamics have been studied by, among others, Schuster and Sigmund (1981), Ritzberger and Weibull (1995), Gaunersdorfer and Hofbauer (1995) and Plank (1997). We summarise the results here.

Theorem 25 (Ritzberger and Weibull 1995) *The semiflow induced by the replicator dynamics (1.9) is volume preserving in the interior of the strategy space for all games. Consequently, a fixed point $\hat{\pi}$ is asymptotically stable if and only if it is a strict Nash equilibrium.*

This result shows that for any game without strict Nash equilibria there will be no asymptotically stable points. Hofbauer (1996) has examined the Hamiltonian nature of this dynamic in two player games, although we do not look at this here.

Theorem 26 (Hofbauer and Weibull 1996) *Actions which are eliminated by the procedure of iterated strict dominance will have zero probability in the limit as $t \to \infty$ for the replicator dynamics (1.9).*

For 2×2 games, the situation is fully analysed by Schuster and Sigmund (1981). In essence, either the game is solvable by iterated strict dominance, in which case

the previous result shows that the Nash equilibrium is asymptotically stable, or there are two pure strategy strict Nash equilibria which are asymptotically stable and an internal equilibrium which is unstable, or there is a unique equilibrium in the interior of the strategy space. For this latter case the replicator admits a constant of motion, and the strategies follow contours of this constant of motion as they cycle around the equilibrium.

For general 2 player games we have the following:

Theorem 27 (Hofbauer and Sigmund 1998) *If the limit set of an orbit of the replicator dynamics (1.9) for 2 player games is contained in the interior of the strategy space then the time average exists and corresponds to a Nash equilibrium.*

On the other hand Plank (1997) gives an example of a three player binary game (where each player has only two actions) for which interior orbits exist with time average not equal to a Nash equilibrium. Little else is known about general N-player games.

We would like to apply Theorem 22 to show that there is zero probability of convergence to a pure strategy which is not a Nash equilibrium. We start by showing that these strategy combinations are linearly unstable.

Proposition 28 *Let π be a fixed point of the replicator dynamics (1.9) which is not a Nash equilibrium. Then π is linearly unstable.*

PROOF Since π is a fixed point, for each i we must have $r^i(a^i, \pi^{-i}) = r^i(\pi)$ whenever $\pi^i(a^i) > 0$. If π is not a Nash equilibrium then, by definition, there exists i, b^i such that $r^i(b^i, \pi^{-i}) > r^i(\pi)$. So we see that $\pi^i(b^i) = 0$. Consider the slightly perturbed joint strategy $\hat{\pi}$ in which player i plays b^i with probability ϵ and chooses from distribution π^i with probability $1 - \epsilon$, for small ϵ, and all other

players play with strategy π^{-i}. Then

$$\begin{aligned}\dot{\hat{\pi}}^i(b^i) &= \hat{\pi}^i(b^i)\left\{r^i(b^i,\hat{\pi}^{-i}) - r^i(\hat{\pi})\right\}\\ &= \epsilon\left\{r^i(b^i,\pi^{-i}) - \epsilon r^i(b^i,\pi^{-i}) - (1-\epsilon)r^i(\pi)\right\}\\ &= \epsilon(1-\epsilon)\left\{r^i(b^i,\pi^{-i}) - r^i(\pi)\right\}\\ &> 0\end{aligned}$$

and so π is linearly unstable.

However if π is a boundary point then the noise is no longer "sufficiently diffuse" to apply Theorem 22. This is apparent from experimental results in Section 2.3, where convergence to non-Nash pure strategy combinations occurs. However, it is still true that there is probability 0 of converging to a completely mixed strategy (i.e. one in the interior of Δ) which is linearly unstable.

Partnership games

Despite the inconclusive nature of the results so far, partnership games are a special case which are more amenable to study. We generalise a result from Hofbauer and Sigmund (1998).

Theorem 20 *In a partnership game the replicator dynamics admit a Lyapunov function for the set of stationary points.*

Proof We drop superscript i on the (expected) reward, since it is identical for all players. Now consider

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}r(\pi) &= \sum_i \sum_{a^i\in A^i} \frac{\partial r}{\partial \pi^i(a^i)}\dot{\pi}^i(a^i)\\ &= \sum_i \sum_{a^i\in A^i} r(a^i,\pi^{-i})\pi^i(a^i)\left(r(a^i,\pi^{-i}) - r(\pi)\right).\end{aligned}$$

But since $\sum_{a^i\in A^i}\pi^i(a^i) = 1$ we see that $\sum_{a^i\in A^i} r(\pi)\pi^i(a^i)\left(r(a^i,\pi^{-i}) - r(\pi)\right) = 0$ and so subtracting from the expression for $\frac{\mathrm{d}}{\mathrm{d}t}r(\pi)$ we see that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}r(\pi) = \sum_i \sum_{a^i\in A^i} \pi^i(a^i)\left(r(a^i,\pi^{-i}) - r(\pi)\right)^2 \geq 0.$$

Equality will hold here only at stationary points of the replicator dynamics, and $-r$ is clearly a Lyapunov function.

Corollary 30 *In partnership games, the learning algorithm (2.1) will converge with probability 1 to the set of fixed points of the replicator dynamics.*

PROOF From Theorems 24 and 29, and Proposition 21, it follows that the limit set of the learning process will consist only of fixed points of the replicator dynamics if the set of possible rewards at these points has empty interior. This fact follows from the Morse–Sard theorem applied to each face of the space of strategies Δ.

2.3 Some examples

In this section we consider our algorithm more carefully in two simple games. For each experiment we use learning parameters $\lambda_n = (n + 100)^{-\rho}$ where the learning rate satisfies $0.5 < \rho \leq 1$. For these 2×2 games, we denote by p the probability that player 1 plays a "head" (i.e. action 1), and by q the probability that player 2 plays a "head". The state of the algorithm is then fully encoded by the pair (p, q).

2.3.1 Simple coordination

We start with a very simple game, which has payoff matrix

$$\begin{pmatrix} (1,1) & (0,0) \\ (0,0) & (0.5,0.5) \end{pmatrix}$$

The fixed points of the replicator dynamics (1.9) for this game are the four pure-strategy combinations, of which only two ($p = q = 1$ and $p = q = 0$) are Nash equilibria, and the interior Nash equilibrium where $p = q = 1/3$. Since the interior equilibrium is linearly unstable, we know from Theorem 22 and Corollary 30 that the learning algorithm will converge to one of the pure-strategy combinations.

It would seem reasonable to expect that our learning algorithm will converge to a random pure strategy equilibria, and will be more likely to converge to the Pareto dominant equilibrium where both players receive reward 1. We ran 1000 simulations, each for 10^5 iterations with $\rho = 0.8$ and a random start point. On 713 of these trials the algorithm "converged" to the Pareto dominant equilibrium, where $p = q = 1$, and on the further 287 trials the algorithm "converged" to the other pure-strategy equilibrium. I have placed the word "converged" in quotation marks here, because obviously the algorithm does not properly converge in a finite number of iterations. However in this situation if the probabilities are within 10^{-3} of either 0 or 1 after 10^5 iterations then we assume the convergence to the relevant point is sufficiently likely.

We can compare the ratios of the number of times to converge to each equilibrium with the size of the basin of the attraction of that equilibrium under the replicator dynamics. From Schuster and Sigmund (1981) we know that these basins have common boundary defined by $p(1-p)^2 = q(1-q)^2$. This clearly has a solution $p = q$, but this is not the relevant line. It is simple algebraic manipulation to see that the boundary line is actually given by $q = \frac{2-p-\sqrt{p(4-3p)}}{2}$. We integrate this from 0 to 1 in to calculate the size of the basin of attraction of the point $(0,0)$.

$$\begin{aligned}
\int_0^1 1 - \frac{p}{2} - \frac{1}{2}\sqrt{p(4-3p)}\,\mathrm{d}p &= \frac{3}{4} - \frac{1}{\sqrt{3}}\int_0^1 \sqrt{1-(\frac{3}{2}p-1)^2}\,\mathrm{d}p \\
&= \frac{3}{4} - \frac{2}{3\sqrt{3}}\int_{-\frac{\pi}{2}}^{\frac{\pi}{6}} \cos^2\theta\,\mathrm{d}\theta \\
&= \frac{2}{3} - \frac{2\sqrt{3}}{27}\pi \\
&\approx 0.2636
\end{aligned}$$

So we see that the proportion of learning episodes that converge to $(0,0)$ compares with the size of the basin of attraction.

2.3.2 2-player matching pennies

This game is the canonical example of a 2×2 game with diametrically opposing objectives for the two players; player 1 scores a point if both players play the same action and player 2 scores a point if the players play opposing actions. The payoff matrix is therefore:

$$\begin{pmatrix} (1,0) & (0,1) \\ (0,1) & (1,0) \end{pmatrix}$$

In this game there is a constant of motion for the replicator dynamics, given by $h(p,q) = p(1-p)q(1-q)$. This constant of motion takes values between 0 (on the boundary of the unit square) and 0.0625 (at the unique equilibrium point where each player plays "head" with probability $\frac{1}{2}$).

Defining $h_n = h(p_n, q_n)$, some simple algebra shows that

$$\mathbb{E}[h_{n+1} \mid p_n, q_n, \lambda_{n+1}] = (1 - \lambda_{n+1}^2)h_n. \tag{2.2}$$

So $\{h_n\}$ is a bounded supermartingale, and thus converges to a random variable h_∞. This explains why our algorithm can converge to an internal cycle instead of a pure strategy equilibrium in this game: for $h_n \to h_\infty$ we must have $\mathbb{E}(h_{n+1} - h_n) \to 0$. If λ_n is bounded away from zero (which of course includes the case of constant λ used by Börgers and Sarin (1997)) then $h_\infty \equiv 0$. On the other hand if $\lambda_n \to 0$ then h_∞ can take values other than 0. This idea is a stochastic analogue of the reasoning used by Akin and Losert (1984) to show that using a simple discretisation of the replicator dynamics in this game will necessarily lead to the boundary.

In fact we can learn more from this analysis: observe that

$$\mathbb{E}[h_\infty \mid h_0] = h_0 \prod_{n=1}^{\infty}(1 - \lambda_n^2). \tag{2.3}$$

If $\sum_{n\geq 1} {\lambda_n}^2 = \infty$ then it follows that $\prod_{n\geq 1}(1 + {\lambda_n}^2) = \infty$. But for $\lambda > 0$ it is clear that $1 - \lambda^2 < (1 + \lambda^2)^{-1}$ and so

$$\prod_{n\geq 1}(1 - {\lambda_n}^2) \leq \prod_{n\geq 1}(1 + {\lambda_n}^2)^{-1} = 0.$$

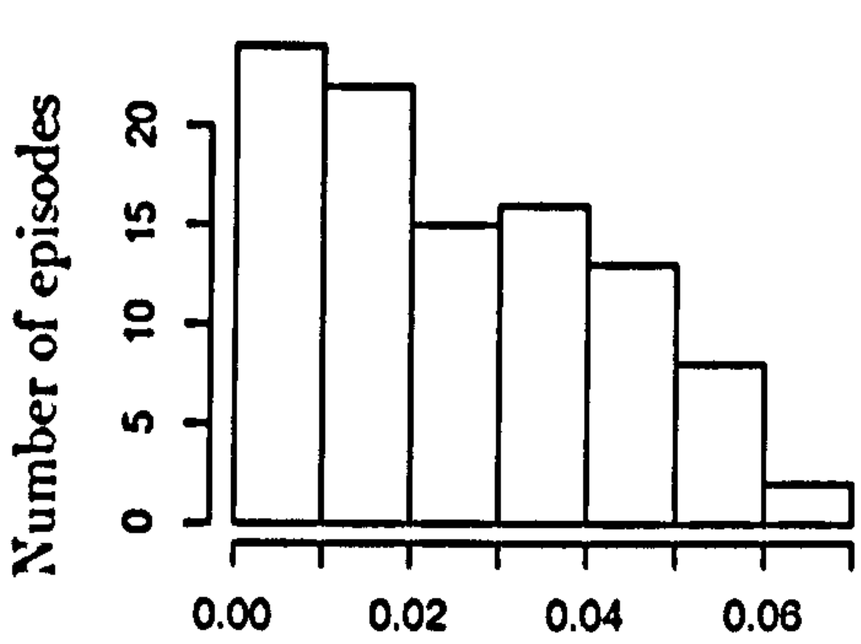


Figure 2.1: Final values of h after 10^5 iterations.

Thus the standard stochastic learning condition of $\sum_{n\geq 1} {\lambda_n}^2 < \infty$ is necessary to avoid the situation of $h_\infty \equiv 0$.

For this game three sets of 100 learning episodes with random start points were run, each with 10^5 iterations for each learning episode. The learning rate ρ was chosen differently for each set. The final value of h was recorded in each case, and histograms of the results are in Figure 2.1. It can be seen that h will tend to finish lower whenever λ decreases more slowly. This is to be expected from equation (2.2), since when λ decreases more slowly the overall drift of h will be larger than

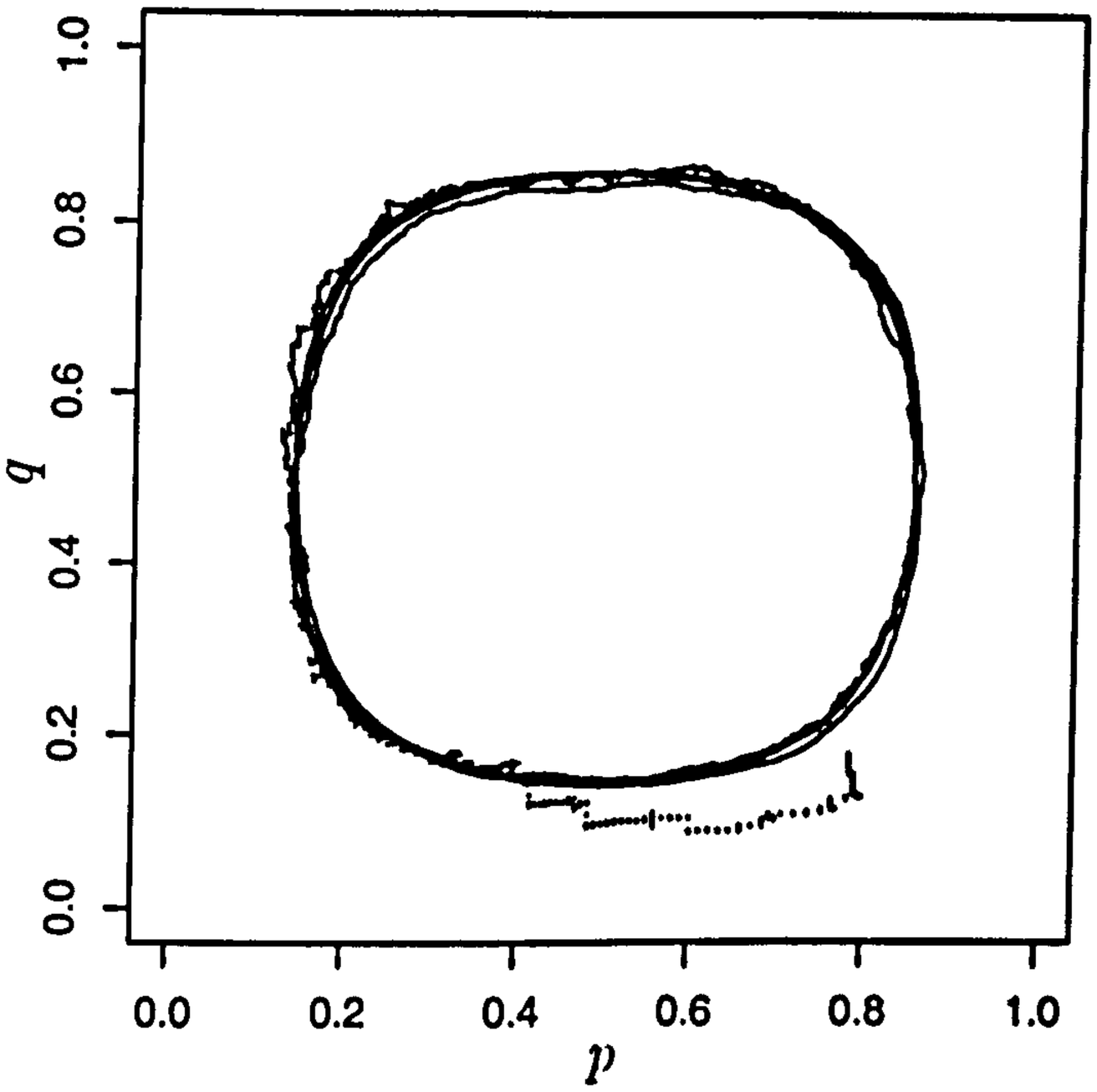


Figure 2.2: Learning trajectory of the simple learning model in 2-player matching pennies.

when λ decreases quickly.

To give a greater appreciation of what happens in this game Fig. 2.2 shows the trajectories through (p, q)-space for a typical learning procedure with $\rho = 0.8$ and 5×10^5 iterations. It can be seen that after initial fluctuations the algorithm settles on to a contour of h, although convergence to a point never occurs.

2.4 An extension

In this section, we restrict our attention to games with positive payoffs. Under this assumption, Hofbauer and Sigmund (1998) show that the payoff-normalised replicator dynamics (1.10) are volume-contracting for 2 player games, and that for generic 2×2 games the set of Nash equilibria is globally attracting for the set of completely mixed strategies.

We could clearly achieve this dynamic via our learning algorithm by simply dividing the obtained reward with a calculated value of $r^i(\pi)$. This however contradicts our basic assumptions that players do not know the structure of the game, nor even that they are playing a game. Instead we learn the value of $r^i(\pi)$ using two-timescales stochastic approximation (Section 1.3.3). Thus our learning algorithm becomes:

$$\pi^i_{n+1}(a^i) = \left(1 - \lambda^{(1)}_{n+1}\frac{R^i_n}{S^i_n}\right)\pi^i_n(a^i) + \lambda^{(1)}_{n+1}\frac{R^i_n}{S^i_n}\mathbb{I}_{\{a^i_n = a^i\}}$$

$$S^i_{n+1} = (1 - \lambda^{(2)}_{n+1})S^i_n + \lambda^{(2)}_{n+1}R^i_n$$

with $\{\lambda^{(i)}_n\}_{n\geq 1}$ satisfying the standard conditions (1.15), and also $\lambda^{(1)}_n/\lambda^{(2)}_n \to 0$ as $n \to \infty$.

We see that S^i_n is our estimate of $r^i(\pi)$. Note that

$$\mathbb{E}[\pi^i_{n+1}(a^i) - \pi^i_n(a^i) \,|\, \pi_n, S_n] = \lambda^{(1)}_{n+1}\pi^i_n(a^i)\frac{r^i(a^i, \pi^{-i}_n) - r^i(\pi_n)}{S^i_n}$$

$$\mathbb{E}[S^i_{n+1} - S^i_n \,|\, \pi_n, S_n] = \lambda^{(2)}_{n+1}\left\{r^i(\pi_n) - S^i_n\right\}.$$

Thus defining $F^1(\pi, S)^i(a^i) = \pi^i(a^i)\left\{r^i(a^i, \pi^{-i}) - r^i(\pi)\right\}/S^i$ and $F^2(\pi, S)^i = r^i(\pi) - S^i$ we can use Theorem 23 (condition **B4** is satisfied since the implicitly defined $U^{(i)}_n$ are martingale differences). The fast ODE (1.21) is simply

$$\dot{S}^i = r^i(\pi) - S^i$$

which, for fixed π, clearly has the globally asymptotically stable fixed point $r^i(\pi)$. This is Lipschitz in π, and so we see that $|S^i_n - r^i(\pi_n)| \to 0$ as $n \to \infty$ and that an interpolation of the π_n is an asymptotic pseudotrajectory of the dynamics

$$\dot{\pi} = F^1(\pi, r^i(\pi)).$$

But by the definition of F^1 this is just the payoff-normalised replicator dynamics (1.10).

There is an additional complication with this algorithm: the stated update to π_n does not necessarily place π_{n+1} into the simplex. In practise this can be resolved

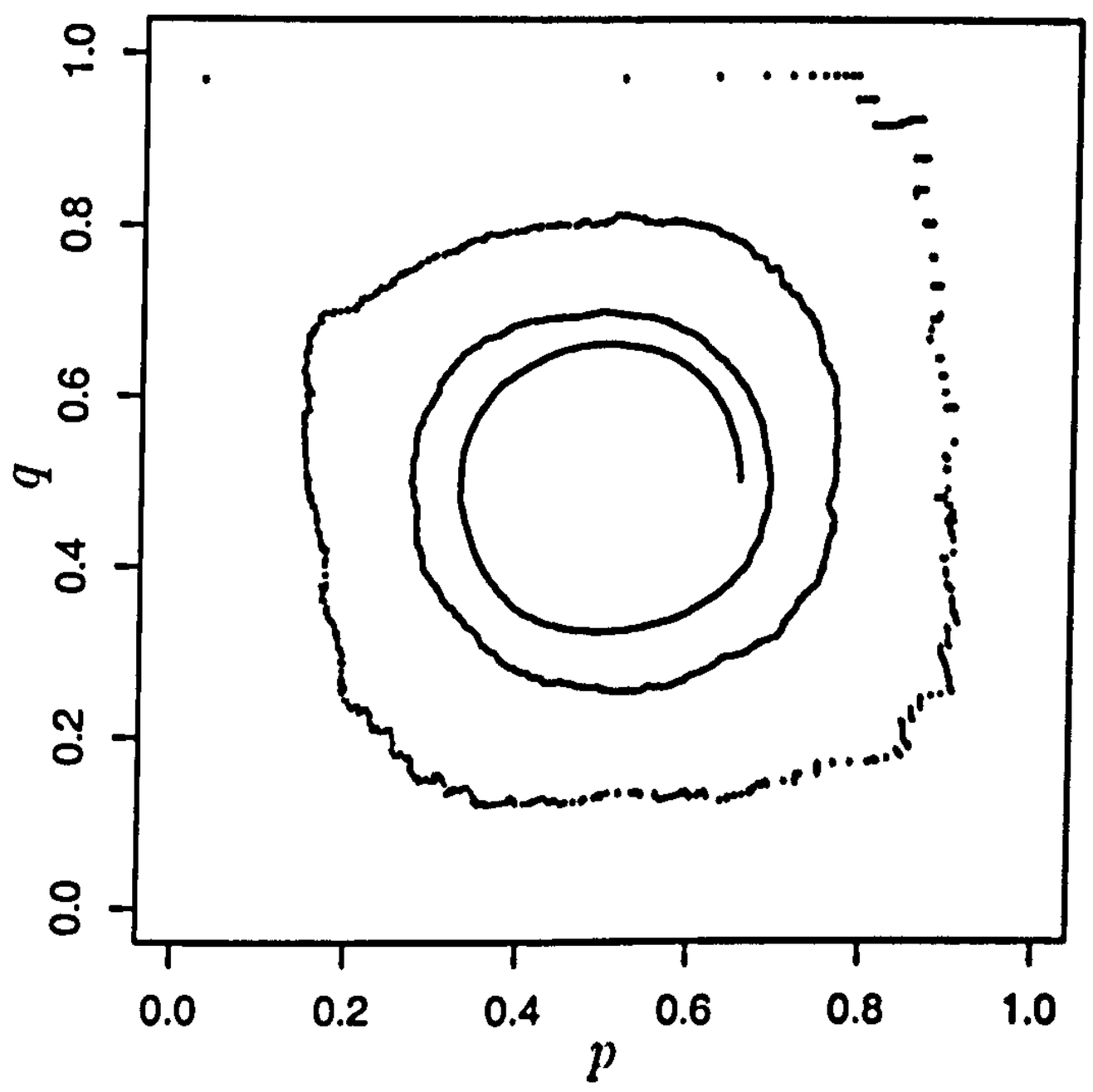


Figure 2.3: Learning trajectory of the normalised learning model in 2-player matching pennies.

by starting λ sufficiently small so that it is not an issue. A further option would be simply to ignore updates that would move π_{n+1} out of the simplex, on the grounds that asymptotically all updates would be included due to the decreasing λ_n.

Convergence of this algorithm is slow. This is partly because as the algorithm nears the equilibrium point the dynamical system (1.10) approximates the un-normalised replicator dynamics (1.9), which does not converge as we have seen. However using $\lambda_n^{(1)} = (n + 100)^{-0.9}$, $\lambda_n^{(2)} = (n + 100)^{-0.6}$, and examining 100 episodes of 5×10^7 iterations, each episode ended in "convergence" to the Nash equilibrium on the matching pennies game which foiled our original algorithm (here "convergence" is even looser than before with convergence being assumed if $h > 0.06$ at the end of the episode). A sample learning episode with 5×10^5 iterations is plotted in Figure 2.3, as was done for the un-normalised learning algorithm in Figure 2.2. It is clear that this algorithm does spiral clockwise towards the centre point, but converges very slowly indeed.

REMARK It will be noted that this is related to the "Win or Learn Fast" (or WoLF) principle (Bowling and Veloso 2002). When a player is doing badly ($S_n^i \approx r^i(\pi_n)$ is small) the adjustments made to π_n^i are greater than if the player is doing well (S_n^i is large).

2.5 Conclusion

We have presented a modification of Börgers and Sarin's learning model, then used standard results of stochastic approximation theory to examine the asymptotic behaviour of our algorithm. We find that the limit set of the learning algorithm is a connected, compact, internally chain-recurrent, invariant set of the semiflow induced by the asymmetric replicator dynamics (1.9).

However we have shown that convergence to a Nash equilibrium is not the only limiting behaviour. Indeed for 2×2 games with a unique internal equilibrium persistent cycling is the norm, and generally it is difficult to rule out convergence

to pure strategy combinations that are not Nash equilibria.

We have also provided a further modification of the algorithm which results in an asymptotic pseudotrajectory of the payoff-normalised replicator dynamics. Although convergence to the boundary of the space of mixed strategies can still not be ruled out, at least limit cycle behaviour is not possible, due to the volume-contracting properties of the semiflow.

This simple model uses observed rewards to directly modify the strategies, when in reality these quantities exist in different spaces. This is addressed in the next chapter where a separate value function is maintained alongside the policy.

Chapter 3

Smooth actor–critic algorithms

The replicator dynamics do not provide a particularly useful framework on which to build a reinforcement learning algorithm, since Nash equilibria are not generally attracting under these dynamics, and there are fixed points that are not Nash equilibria. A more promising dynamical system is based upon best responses. Such a reinforcement learning algorithm will then relate closely to fictitious play and its variants.

Several parts of this chapter will appear in Leslie and Collins (2003), and were presented at the 13th International Conference on Game Theory at Stonybrook.

3.1 Introduction

As observed in Section 1.1.3, the beliefs under fictitious play will converge to Nash equilibrium in several classes of games. Suppose that, instead of maintaining beliefs σ_n^i about opponent strategies, players are told the current opponent strategy π_n^{-i} and adjust their own strategy towards a best response to this. This results in a best-response adaptation procedure

$$\pi_{n+1}^i = (1 - \lambda_{n+1})\pi_n^i + \lambda_{n+1}\mathrm{BR}(\pi_n^{-i}) \tag{3.1}$$

for some learning parameters $\{\lambda_n\}_{n\geq 1}$. It is immediately obvious that if $\lambda_n = n^{-1}$ then the π_n evolve exactly as the beliefs σ_n of a fictitious play process (1.6). However, if this scheme were realisable it would be the strategies of the players that converges to equilibrium, as opposed to empirical averages.

At first, it seems that we have taken a step backwards here, in that now players need to observe the full mixed strategy of the opponents, whereas for fictitious play all that is required is the action played at each step. However, note from the definition (1.1) of best responses that all that is required to select a best response is a knowledge of the expected rewards $r^i(a^i, \pi^{-i})$. In this chapter we will show how to use two-timescales stochastic approximation to learn these expected rewards, without any observation of opponent play being required at all.

However, in order to use the ODE method of stochastic approximation it is necessary for the expected adjustment to be a Lipschitz function of the current strategy; best responses are not continuous in opponent strategy. Also, as we have seen in Chapter 2, if players approach a pure strategy (or a mixed strategy where any action is played with zero probability) then the player may never realise that unplayed actions are better responses to opponent strategies. A similar problem arises here. Therefore, following Fudenberg and Kreps (1993), we consider not best responses (1.1) but smooth best responses (1.5). In Chapter 6 we will see how to adapt this algorithm to use best responses.

The chapter is arranged as follows. In the next section we discuss stochastic fictitious play (Fudenberg and Kreps 1993; Benaïm and Hirsch 1999) and the smooth best response dynamics (Hofbauer and Hopkins 2000). We present and analyse our algorithm in Section 3.3, then demonstrate a version for symmetric games (similar to the symmetric stochastic fictitious play of Hofbauer and Sandholm (2002)) in Section 3.4. A numerical example is presented in Section 3.5.

3.2 Stochastic fictitious play

Fudenberg and Kreps (1993) observe that convergence of beliefs in a fictitious play process does not mean that strategies, or even average payoffs, converge to Nash equilibrium. To counter this they introduced stochastic fictitious play. Here, at time n, player i plays the mixed strategy corresponding to a smooth best response to the beliefs σ_n^{-i} about opponent strategy. The beliefs at time n are the average of the previously observed actions. Thus

$$\mathbb{E}[\sigma_{n+1} - \sigma_n \,|\, \sigma_n] = \tfrac{1}{n+1}\,(\beta(\sigma_n) - \sigma_n)$$

Since β is Lipschitz, the strategies are bounded (they remain in Δ), and $\sigma_{n+1} = \sigma_n + \lambda_{n+1}\{\beta(\sigma_n) - \sigma_n + U_{n+1}\}$ where the U_n are bounded martingale differences, the following is immediate:

Theorem 31 (Benaïm and Hirsch 1999) *A suitable interpolation of the sequence $\{\sigma_n\}_{n\geq 1}$ of beliefs arising from a stochastic fictitious play process is an asymptotic pseudotrajectory of the semiflow induced by the smooth best response dynamics (1.8).*

Thus we are interested in the chain-recurrent sets of the smooth best response dynamics. The relevant results are given by Hofbauer and Sandholm (2002).

Theorem 32 *The smooth best response dynamics (1.8) admit a Lyapunov function for the set of Nash distributions in the cases of 2-player zero-sum games and of N-player partnership games. For 2-player zero-sum games there is a unique Nash distribution (for any given smooth best response functions).*

PROOF Consider first partnership games, and the function

$$U(\pi) = r(\pi) + \tau \sum_{i=1}^{N} v^i(\pi^i),$$

where $r(\cdot)$ denotes the (common) reward function, τ denotes the temperature parameter, and v^i is the smoothing function used by player i to calculate the smooth

best response β^i. Note that, by the definition (1.5) of smooth best responses, $\{r^i(\cdot,\pi^{-i}) + \tau\nabla v^i(\beta^i(\pi^{-i}))\}\cdot\xi = 0$ for any ξ in the tangent space to Δ^i, and in particular for $\xi = \dot\pi^i$. So

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}U(\pi) &= \sum_i \nabla_{\pi_i}U(\pi)\cdot\dot\pi^i\\ &= \sum_i \left(r(\cdot,\pi^{-i}) + \tau\nabla v^i(\pi^i)\right)\cdot\left(\beta^i(\pi^{-i}) - \pi^i\right)\\ &= \tau\sum_i \left(-\nabla v^i(\beta^i(\pi^{-i})) + \nabla v^i(\pi^i)\right)\cdot\left(\beta^i(\pi^{-i}) - \pi^i\right).\end{aligned}$$

By the strict concavity of the v^i, $\frac{\mathrm{d}}{\mathrm{d}t}U(\pi) \geq 0$ with equality only when $\pi^i = \beta^i(\pi^{-i})$ for all i, i.e. when π is a Nash distribution. Thus $-U$ is a Lyapunov function.

For 2-player zero-sum games, define functions

$$V^i(\pi) = r^i(\beta^i(\pi^{-i}),\pi^{-i}) + \tau v^i(\beta^i(\pi^{-i})) - r^i(\pi) - \tau v^i(\pi^i)$$

for $i = 1,2$. Then

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}V^i(\pi) = {}& \left\{r^i(\cdot,\pi^{-i}) + \tau\nabla v^i(\beta^i(\pi^{-i}))\right\}\cdot\tfrac{\mathrm{d}}{\mathrm{d}t}\beta^i(\pi^{-i})\\ &- \left\{r^i(\cdot,\pi^{-i}) + \tau\nabla v^i(\pi^i)\right\}\cdot\dot\pi^i\\ &+ \left\{r^i(\beta^i(\pi^{-i}),\cdot) - r^i(\pi,\cdot)\right\}\cdot\dot\pi^{-i}.\end{aligned}$$

By the definition of β^i, $\{r^i(\cdot,\pi^{-i}) + \tau\nabla v^i(\beta^i(\pi^{-i}))\}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\beta^i(\pi^{-i}) = 0$, and as before $r^i(\cdot,\pi^{-i})\cdot\dot\pi^i = -\tau\nabla v^i(\beta^i(\pi^{-i}))\cdot\dot\pi^i$. Therefore

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}V^i(\pi) = {}& \tau\left\{\nabla v^i(\beta^i(\pi^{-i})) - \nabla v^i(\pi^i)\right\}\cdot\left(\beta^i(\pi^{-i}) - \pi^i\right)\\ &+ \left\{r^i(\beta^i(\pi^{-i}),\beta^{-i}(\pi^i)) - r^i(\beta^i(\pi^{-i}),\pi^{-i}) - r^i(\pi^i,\beta^{-i}(\pi^i)) + r^i(\pi)\right\}\end{aligned}$$

Thus taking $V = V^1 + V^2$ gives

$$\tfrac{\mathrm{d}}{\mathrm{d}t}V(\pi) = \tau\sum_{i=1,2}\left\{\nabla v^i(\beta^i(\pi^{-i})) - \nabla v^i(\pi^i)\right\}\cdot\left(\beta^i(\pi^{-i}) - \pi^i\right),$$

since the game is zero-sum. Again, by the strict concavity of the v^i, $\frac{\mathrm{d}}{\mathrm{d}t}V(\pi) \leq 0$ with equality only when π is a Nash distribution.

We show that there is a unique Nash distribution for 2-player zero-sum games by showing that V is strictly convex, and so has a unique minimum. Since Nash distributions are necessarily minima of V the result will follow. Write

$$\begin{aligned} V(\pi) =& V^1(\pi) + V^2(\pi) \\ =& r^1(\beta^1(\pi^2), \pi^2) + \tau v^1(\beta^1(\pi^2)) - \tau v^2(\pi^2) \\ &+ r^2(\beta^2(\pi^1), \pi^1) + \tau v^2(\beta^2(\pi^1)) - \tau v^1(\pi^1) \end{aligned}$$

and define $W^i(\pi^{-i}) = r^i(\beta^i(\pi^{-i}), \pi^{-i}) + \tau\{v^i(\beta^i(\pi^{-i})) - v^{-i}(\pi^{-i})\}$, so that $V(\pi) = W^1(\pi^2) + W^2(\pi^1)$. Now $r^i(\beta^i(\pi^{-i}), \pi^{-i}) + \tau v^i(\beta^i(\pi^{-i}))$ is convex, since it is the maximum over linear functions of π^{-i}, and $v^{-i}(\pi^{-i})$ is strictly concave, so $W^i(\pi^{-i})$ is strictly convex in π^{-i}. Therefore $V(\pi)$ is strictly convex in π, being the sum of two strictly convex functions of independent components of π.

REMARK Although Hofbauer and Hopkins (2000) claim that V is convex, their proof is incomplete and does not take into the account the zero-sum character of the game. Thus, if correct, their proof would show that all games have a unique Nash distribution (which is not true—consider a simple 2×2 coordination game where each player gets a point when they play the same action and scores nothing otherwise).

Corollary 33 *Under a stochastic fictitious play process in a 2-player zero-sum game, the strategies of the players will converge to the unique Nash distribution. In an N-player partnership game with finite or countably many Nash distributions, the strategies will converge to one of these Nash distributions.*

PROOF For a 2-player zero-sum game, the beliefs will converge with probability 1 to the unique Nash distribution. This follows immediately from Theorem 31, Proposition 21, and Theorem 32. Similarly, for an N-player partnership game the beliefs will converge with probability 1 to a connected set of Nash distributions on which the Lyapunov function is constant. But in both cases the strategies

employed by the players are continuous functions of the beliefs, and the result follows by the definition of Nash distributions.

REMARK Note that the strategies played, i.e. $\beta^i(\pi^{-i})$, do not necessarily follow the best response dynamics at all: Sato and Crutchfield (2002) give evidence to suggest that the strategies can follow a modified replicator dynamics.

There are two classical examples of games for which classical fictitious play (1.6) fails to converge to equilibrium. Shapley (1964) proposed a modification of a rock–scissors–paper game, equivalent to the following:

$$\begin{pmatrix} (0,0) & (1,0) & (0,1) \\ (0,1) & (0,0) & (1,0) \\ (1,0) & (0,1) & (0,0) \end{pmatrix} \tag{3.2}$$

Jordan (1993) proposed a 3-player matching pennies game where each player can choose either heads or tails. Player 1 scores a point if they match player 2, player 2 scores a point if they match player 3, and player 3 scores a point if they choose the opposite to player 1. The payoff 'matrix' is given by

$$\begin{array}{lcc} \text{Player 3's choice:} & \text{HEAD} & \text{TAIL} \\ \text{Player 2's choice:} & \text{HEAD}\quad\text{TAIL} & \text{HEAD}\quad\text{TAIL} \\ \text{Player 1's choice} \begin{cases} H \\ T \end{cases} & \begin{pmatrix} (1,1,0) & (0,0,0) \\ (0,1,1) & (1,0,1) \end{pmatrix} & \begin{pmatrix} (1,0,1) & (0,1,1) \\ (0,0,0) & (1,1,0) \end{pmatrix} \end{array} \tag{3.3}$$

where an entry (r^1, r^2, r^3) gives the rewards to players 1, 2 and 3 respectively. For both of these games it has been shown that the unique Nash distribution is linearly unstable for the smooth best response dynamics with sufficiently small temperature $\tau > 0$—Cowan (1992) showed this for Shapley's game, while Benaïm and Hirsch (1999) give the result for Jordan's game. Using Pemantle's result (Theorem 22) shows that convergence to the Nash distribution happens with probability zero for these games, and in fact an attracting limit cycle is present in both cases.

3.3 A two-timescales learning algorithm

We now present an algorithm that also results in an asymptotic pseudotrajectory to the smooth best response dynamics, but which does not require players to observe the actions of others, or to know the structure of the game. As already noted, the only reason players need this information is so that they can estimate the expected value of each of their actions in order to calculate the smooth best response. Reinforcement learning is a model-free alternative for estimating expected values of a set of actions, although it relies on the fact that these expected values do not change with time.

Assume we have a stationary random environment where at each stage player i must choose an action a^i from a finite set A^i, and associated with each action $a^i \in A^i$ there is a random reward $R(a^i)$ which has a fixed distribution and bounded variation. Consider the learning scheme

$$Q_{n+1}(a^i) = Q_n(a^i) + \lambda_{n+1}\mathbb{I}_{\{a_n^i = a^i\}}(R_n^i - Q_n(a^i)), \quad \text{for } a^i \in A^i,$$

where a_n^i is the action chosen at stage n, R_n^i is the subsequent reward, and $\{\lambda_n\}_{n\geq 1}$ is a deterministic sequence satisfying the standard conditions (1.15). It is well-known in the reinforcement learning literature (Sutton and Barto 1998; Bertsekas and Tsitsiklis 1996) that, provided each action is chosen infinitely often, the Q values in this algorithm will converge almost surely to the expected action values, i.e.

$$Q_n(a^i) \to \mathbb{E}[R(a^i)] \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

However in our multi-agent setting the players' strategies are all changing simultaneously as each player learns, and consequently the sampled rewards, R_n^i, do not come from a stationary distribution. So when learning $r^i(a^i, \pi^{-i})$ the standard results no longer apply. A solution is to be found in Borkar's two-timescales stochastic approximation (Borkar 1997), inspired by similar techniques used by Konda and Borkar (2000) and Borkar (2001). Intuitively, if players update their

Q values on a fast timescale, while adjusting strategies on a slow timescale, then the Q values will give asymptotically accurate estimates of $r^i(a^i, \pi^{-i})$. Players use these asymptotically accurate estimates to adjust their strategies according to a smoothed version of the best-response adaptation procedure (3.1).

Because players no longer play smooth best responses to strategies, but instead to perceived values, we will write

$$\beta^i(Q^i) = \operatorname*{argmax}_{\pi^i \in \Delta^i} \left\{ \sum_{a^i \in A^i} \pi^i(a^i) Q^i(a^i) + \tau v^i(\pi^i) \right\}. \tag{3.4}$$

Then $\beta^i(r^i(\cdot, \pi^{-i}))$ gives what would be expected (i.e. $\beta^i(\pi^{-i})$ in the previous notation).

Actor–critic algorithm

Each player i selects an action a_n^i using the strategy π_n^i, then updates π_n^i and Q_n^i according to

$$\begin{aligned} \pi_{n+1}^i &= (1 - \mu_{n+1})\pi_n^i + \mu_{n+1}\beta^i(Q_n^i) \\ Q_{n+1}^i(a^i) &= Q_n^i(a^i) + \lambda_{n+1}\mathbb{I}_{\{a_n^i = a^i\}}(R_n^i - Q_n^i(a^i)), \quad \text{for } a^i \in A^i, \end{aligned} \tag{3.5}$$

where $\{\lambda_n\}_{n\geq 1}$ and $\{\mu_n\}_{n\geq 1}$ are deterministic sequences satisfying the standard conditions (1.15), and $\mu_n/\lambda_n \to 0$ as $n \to \infty$.

We consider the Q_n^i process to be a critic process, assessing the value of the current strategy, while the π_n^i is an actor process, adjusting based upon information provided by critic. The conditions placed on $\{\lambda_n\}_{n\geq 1}$ and $\{\mu_n\}_{n\geq 1}$ mean that $Q_n^i(a)$ places greater emphasis on recent observations of $r^i(a, \pi_n^{-i})$ than on observations from the distant past (as opposed to if $\lambda_n \approx 1/n$, in which case all observations would have equal weighting in the calculation of Q_n). This is a sensible thing to do, given that π_n^{-i} is evolving.

Theorem 34 *An interpolation of the strategies π_n played during the actor-critic algorithm (3.5) is an asymptotic pseudotrajectory of the smooth best response dynamics (1.8).*

PROOF Note that the Q values of any player will remain bounded, and so the strategies will stay in a region where $\pi_n^i(a^i)$ is bounded away from 0 and 1. Take these regions as our metric spaces for Theorem 23, and define

$$F^{(1)}(\pi, Q)^i = \beta^i(Q^i) - \pi^i, \quad \text{for each } i,$$

$$F^{(2)}(\pi, Q)^i(a^i) = \pi^i(a^i)\left\{r^i(a^i, \pi^{-i}) - Q^i(a^i)\right\}, \quad \text{for each } i \text{ and } a^i.$$

The conditions **B1–B4** of Theorem 23 are met, and we must analyse the slow and fast ODEs (corresponding to (1.22) and (1.21)). The fast ODE is simply

$$\dot{Q}^i(a^i) = \pi^i(a^i)\left\{r^i(a^i, \pi^{-i}) - Q^i(a^i)\right\}, \quad \text{for each } i \text{ and } a^i,$$

which, for fixed π has a unique globally asymptotically stable fixed point where $Q^i(a^i) = r^i(a^i, \pi^{-i})$ (since our metric space of definition excludes points where $\pi^i(a^i) = 0$). Defining

$$\underline{r}(\pi) = (r^1(\cdot, \pi^{-i}), \ldots, r^N(\cdot, \pi^{-i})), \tag{3.6}$$

the conclusions of Theorem 23 state that, with probability 1, $\|Q_n - \underline{r}(\pi_n)\|_\infty \to 0$ as $n \to \infty$ and an interpolation of the π_n is an asymptotic pseudotrajectory of the semiflow induced by

$$\dot{\pi} = F^{(1)}(\pi, \underline{r}(\pi)), \quad \text{or equivalently}$$

$$\dot{\pi}^i = \beta^i(r^i(\cdot, \pi^{-i})) - \pi^i \quad \text{for each } i.$$

But this is simply the smooth best response dynamics (1.8).

Thus we see that the actor–critic algorithm (3.5) is asymptotically equivalent to the stochastic fictitious play algorithm, in that both result in asymptotic pseudo-trajectories of the smooth best response dynamics.

Corollary 35 *Under the actor–critic algorithm (3.5) in a 2-player zero-sum game, the strategies of the players will converge to the unique Nash distribution. In an N-player partnership game with finite or countably many Nash distributions, the strategies will converge to one of these Nash distributions.*

PROOF This proof is exactly the same as that of Corollary 33.

However, the instability of the Nash distribution of Shapley's game (3.2) and of Jordan's game (3.3) under the smooth best response dynamics still presents difficulties. While it seems reasonable that an analogous non-convergence result to Theorem 22 should hold in this case, this noise is only present on the fast timescale; given (π_n^i, Q_n^i) the update to π_n^i is deterministic, and so the probabilistic estimates used by Pemantle are not valid in this case. However the presence of an attracting orbit in each case means that, by a simple extension of the results of Benaïm (1999) discussed in Section 1.3, the probability of convergence to the equilibrium is less then 1.

REMARK We could modify our algorithm so that at time n player i selects a random action to reinforce, using $\beta^i(Q_n^i)$ to select this action. Then the update to π_n^i will be random, as in stochastic fictitious play, and Theorem 22 will hold. However, it is not clear that introducing this extra level of randomness will improve the convergence properties of the algorithm, and indeed it seems likely that a slightly more sophisticated approach to the problem of non-convergence to unstable fixed points will provide the required result.

Despite these non-convergence results, the following is true:

Theorem 36 *If the actor–critic algorithm (3.5) converges to a fixed point*

$$(Q_n, \pi_n) \to (\tilde{Q}, \tilde{\pi}) \quad as \quad n \to \infty$$

then $\tilde{Q}^i(a^i) = r^i(a^i, \tilde{\pi}^{-i})$ *and* $\tilde{\pi}$ *is a Nash distribution.*

PROOF It is a basic result of stochastic approximation theory, and essentially a law of large numbers, that if convergence occurs then the limit point must be a zero of the associated ODE. It follows immediately that $\tilde{Q}^i(a^i) = r^i(a^i, \tilde{\pi}^{-i})$, and $\beta^i(\tilde{Q}^i) = \tilde{\pi}^i$. Therefore $\tilde{\pi}^i = \beta^i(\tilde{\pi}^{-i})$.

3.4 Symmetric games

We can study a very similar process in symmetric games. Essentially we assume that a population maintains a set of Q values, while adapting more slowly than the Q values are updated. There are (at least) two possible motivations for this model:

Model 1: Each individual of an infinite population plays a pure strategy. The population state at time n is π_n. A random individual, who uses action a_n, is selected to play the game, and receives reward $r(a_n, \pi_n)$. Each member of the population observes this, and adjusts $Q(a_n)$ towards the newly observed value using

$$Q_{n+1}(a_n) = (1 - \lambda_{n+1})Q_n(a_n) + \lambda_{n+1} r(a, \pi_n).$$

Also, a fraction μ_n of the population will switch their strategy, selecting a new strategy using a smooth best response to the current estimates Q_n. This is similar to a stochastic fictitious play interpretation used by Hopkins (1999).

Model 2: Players join a game at distinct time points, choosing an action as they join and using that action for all time. As a player enters the game, they receive a reward depending on their choice of action and the state of the population already playing the game. Players not yet participating in the game observe the rewards of players as they enter, and maintain Q values in the usual fashion. When a player enters the game, they select an action using a smooth best response to the current Q values. This is similar to the symmetric stochastic fictitious play model of Hofbauer and Sandholm (2002).

In each case, the Q values and population states evolve according to

$$\pi_{n+1} = \pi_n + \mu_{n+1} \left\{ \beta(Q_n) - \pi_n + U_n^1 \right\}$$

$$Q_{n+1}(a) = Q_n(a) + \lambda_{n+1} \mathbb{I}_{\{a_n = a\}} \left\{ r(a, \pi_n) - Q_n(a) + U_n^2 \right\} \quad \text{for each } a \in A.$$

In Model 1, the U_n^1 are identically 0, since the infinite population size renders the population adjustment deterministic. In Model 2, $\mu_{n+1} = 1/(n+1)$, since the population state is simply the average of the actions previously selected. A second difference is the (effective) mixed strategy actually played at any particular step. Under Model 1, the player is selected from the population, and so the mixed strategy is effectively π_n. On the other hand, under Model 2 the player joining selects an action according to $\beta(Q_n)$. However, since the reward to the player depends on the population state π_n in both cases, the analysis is identical.

As with our actor–critic learning algorithm (3.5), we take $\{\lambda_n\}_{n\geq 1}$ and $\{\mu_n\}_{n\geq 1}$ to be sequences satisfying the standard conditions (1.15), but also with $\mu_n/\lambda_n \to 0$ as $n \to \infty$. Again, this means that recent observations of $r(a, \pi_n)$ have more influence on $Q_n(a)$ than distant observations.

Proposition 37 *An interpolation of the population states under either Model 1 or Model 2 is an asymptotic pseudotrajectory of the symmetric smooth best response dynamics (1.8).*

PROOF The proof of this proposition is directly analogous to that of Theorem 34.

Hofbauer and Sandholm (2002) study the symmetric smooth best response dynamic, motivated by a study of symmetric stochastic fictitious play. Since stochastic fictitious play also results in asymptotic pseudotrajectories of the smooth best response dynamic, they investigate the chain-recurrent sets. As well as the classes of games we specified in Section 1.1, they consider supermodular games, for which there is an ordering on the actions such that for $a' > a$, $b' > b$ we have $r(a', b') - r(a, b') > r(a', b) - r(a, b)$. That is, when one player moves to a 'higher' action the incentive for the other to also switch to a higher action increases.

Theorem 38 (Hofbauer and Sandholm 2002) *Consider the smooth best response dynamics (1.8). The following is true:*

1. *for zero sum games there is a unique Nash distribution $\tilde{\pi}$, and $\{\tilde{\pi}\}$ is the unique chain-recurrent set,*

2. *for games with an interior ESS there is a unique Nash distribution $\tilde{\pi}$, and $\{\tilde{\pi}\}$ is the unique chain-recurrent set,*

3. *for partnership games with isolated Nash distributions, any connected, internally chain recurrent set is a Nash distribution.*

In fact it is shown by Hofbauer (2000) that the first two classes are special cases of the class of games for which the rewards are given by a matrix $U = (u_{ab})$, with $r(a, b) = u_{ab}$ and

$$\xi^T U \xi \leq 0 \quad \forall \xi \in \mathbb{R}_0^{|A|} = \{\xi \in \mathbb{R}^{|A|} : \sum_i \xi_i = 0\},$$

and that the conclusion for those two classes continues to hold for this wider class.

Theorem 30 *Under Model 1 or Model 2, the population state will converge to a Nash distribution in any of the following classes of games:*

1. *zero-sum games,*

2. *games with an internal ESS,*

3. *partnership games with isolated Nash distributions,*

PROOF For all of these classes, any chain-recurrent set is a Nash distribution, and so by Proposition 37 and Theorem 20 the result follows.

3.5 A numerical example

In this section we illustrate Model 1 using a simple rock–scissors–paper game, with payoff matrix

$$\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}.$$

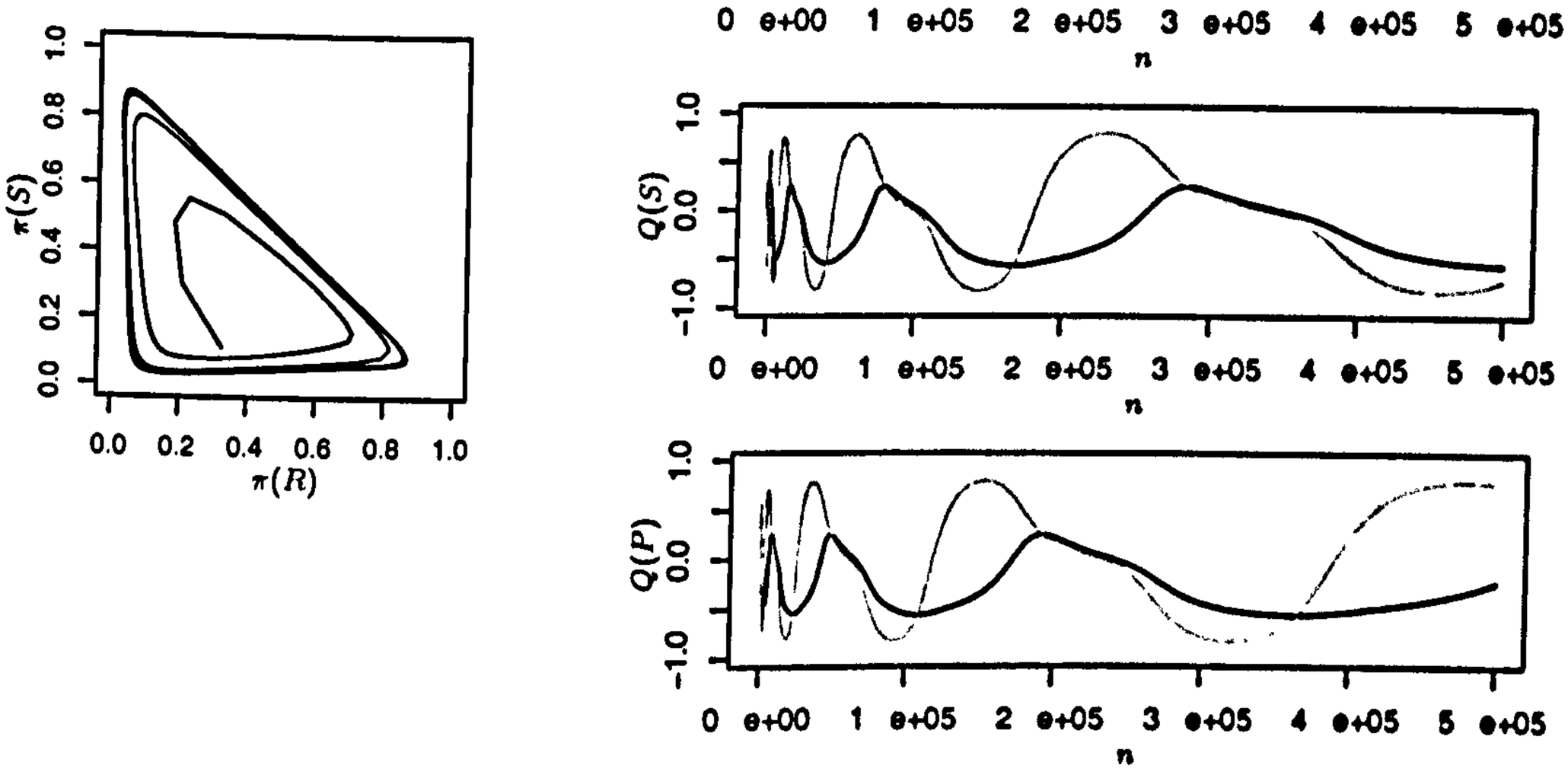


Figure 3.1: Continuous clockwise cycling of strategies (left) and non-convergence of Q values to $r(\cdot, \pi)$ (right), using $\lambda_n = \mu_n = n^{-0.8}$.

We use this model because the dimensions are smaller than with the actor-critic algorithm of Section 3.3 (since there is effectively only one player) and so visualisation is simpler. The experiments will demonstrate the need to use separate learning parameters for Q and π for this game.

Boltzmann smooth best responses (1.16) were used, with temperature parameter $\tau = 0.2$. In each case the experiment was run for 5×10^5 iterations, starting at a random start point. Points were plotted every 50 iterations.

In the first experiment $\lambda_n = \mu_n$ for all n, resulting in a single-timescale stochastic approximation. The results are shown in Fig. 3.1. It is clear from the left hand diagram that the strategies persistently cycle clockwise in this case. Of more interest are the diagrams on the right, comparing the current estimates of the values, Q, with the calculated current value of the actions, $r(\cdot, \pi)$. In each diagram the Q value is plotted in black while the calculated value is in grey. It is clear that the Q values cannot 'keep up' with the calculated values, which adjust at the same

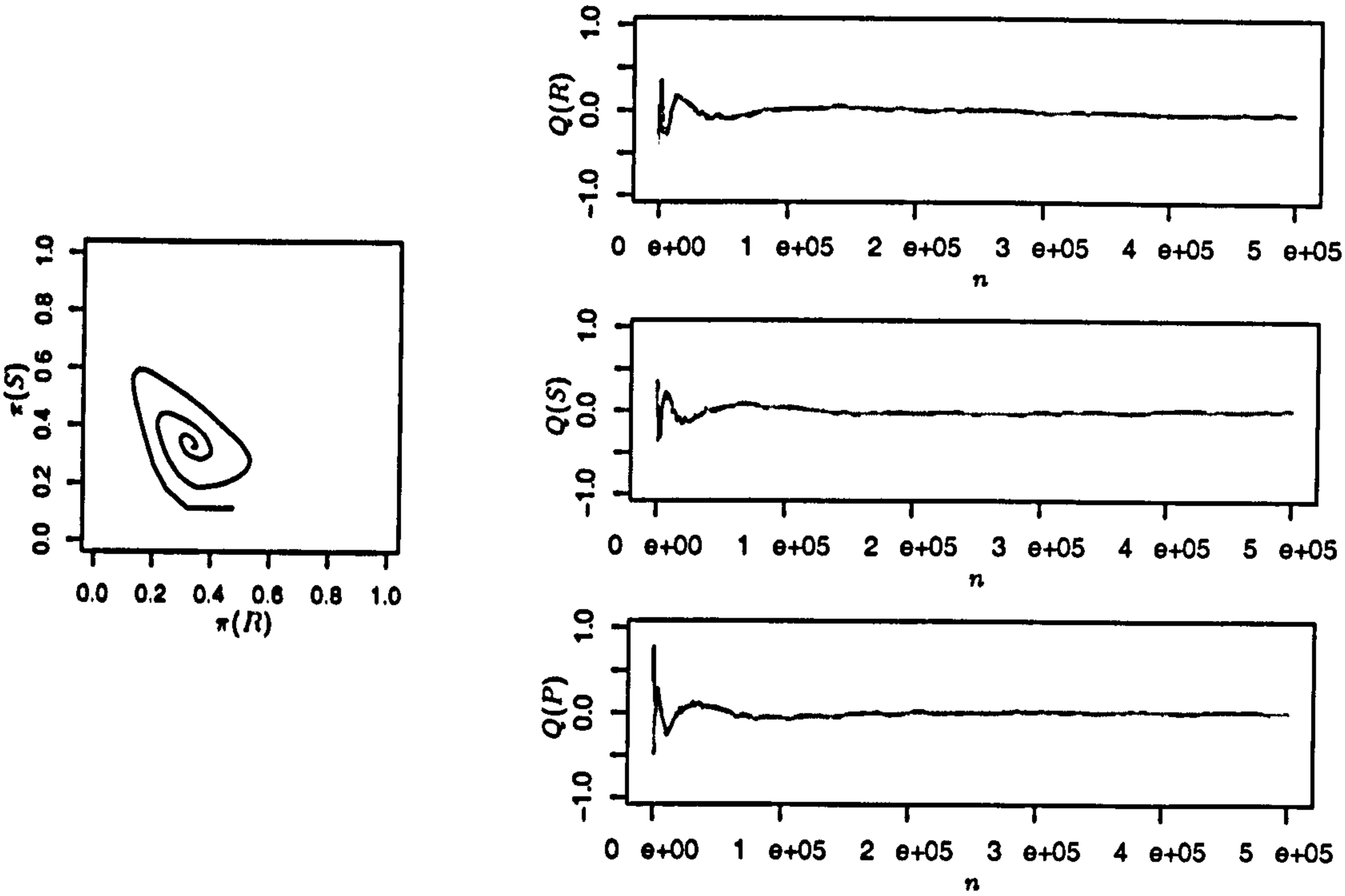


Figure 3.2: Convergence of strategies in a clockwise spiral (left) and convergence of Q values to $r(\cdot, \pi)$ (right), using $\lambda_n = n^{-0.7}$, $\mu_n = n^{-1}$.

rate as the population state π. The 'stretching' of the process as n increases in these plots is due to the fact that as n increases the learning parameters decrease, and so the processes adjust more slowly.

For the second experiment we introduce our second timescale, so that the Q values update on a faster timescale than the population state. There is a significant change in the diagrams. The left hand diagram in Fig. 3.2 shows a convergent trajectory (although after 5×10^5 iterations it has not yet fully converged). Again, of much greater interest is the right hand set of diagrams. In each case it is clear that the Q value tracks the calculated value $r(\cdot, \pi)$ very closely indeed (again the Q value is plotted in black whereas the calculated value is in grey). This corresponds to the result saying that $\|Q_n - A\pi_n\| \to 0$ as $n \to \infty$.

These two experiments suggest that for this particular game the use of two timescales is necessary to allow the Q values to successfully estimate $r(\cdot, \pi)$. For

simpler games, such as the Hawk–Dove game (Hofbauer and Sigmund 1998), without an inherent cycling of the strategies under the smooth best response dynamics, the use of two timescales proves to be unnecessary.

3.6 Conclusion

In contrast with the model of the previous chapter, we explicitly addressed the relationship between rewards and actions, by using separate value functions and policies in an actor–critic algorithm. We used two-timescales stochastic approximation to let players learn the values of actions, and then mapped these value estimates to policy space using the smooth best response function (3.4). The policies adapt towards this smooth best response on the slow timescale, and it follows that the value estimates are asymptotically accurate. This in turn means that our process results in asymptotic pseudotrajectories of the smooth best response dynamics (1.8), and should it converge to a fixed point then that fixed point will be a Nash distribution, as opposed to a Nash equilibrium.

Therefore our process has similar properties to stochastic fictitious play (Fudenberg and Kreps 1993; Benaïm and Hirsch 1999; Hofbauer and Sandholm 2002). By combining results from the literature (Benaïm and Hirsch 1999; Hofbauer and Hopkins 2000; Leslie and Collins 2003) we showed that stochastic fictitious play converges in a larger class of games than has been previously stated (Corollary 33). The same result shows that our actor–critic algorithm must converge to Nash distribution in all the same games that stochastic fictitious play will converge, despite each player using less information—under stochastic fictitious play each player must be able to observe the actions played by the others, and to calculate the value of their actions using this information, while under the actor–critic algorithm each player need only observe their own reward each time they play an action.

Two models of evolution of a population in a symmetric game were also con-

sidered, and again the convergence results are the same as those obtained for symmetric stochastic fictitious play (Hofbauer and Sandholm 2002).

A numerical example was given to show that the separation of learning parameters between the learning of the Q values and the adaptation of the strategies π is necessary to allow Q_n to be an accurate assessment of $r(\cdot, \pi_n)$, and therefore for the strategies to converge.

However, the two classical examples for which fictitious play does not converge (Shapley's game (3.2) and Jordan's game (3.3)) still present problems. It has been shown (Cowan 1992; Benaïm and Hirsch 1999) that a limit cycle is attracting for these games, and so convergence to equilibrium cannot be guaranteed (indeed it seems likely that the probability of the actor–critic algorithm (3.5) converging to equilibrium is zero, but this has not been shown). In order to break the symmetry between the players that causes this cycling behaviour, we will extend the actor–critic learning algorithm in the next chapter.

Chapter 4

Multiple timescales

The non-convergence of the actor–critic algorithm (3.5) in certain games motivates a further extension. Littman and Stone's work (2001) suggests the consideration of 'leaders' and 'followers'; this can be achieved by studying players that learn at different rates. To analyse these players requires an extension of Borkar's result (Theorem 23) beyond two timescales.

Much of this chapter will appear in Leslie and Collins (2003).

4.1 Stochastic approximation with multiple timescales

Consider N interdependent stochastic approximation processes $\theta_n^{(1)}, \ldots, \theta_n^{(N)}$, each lying in a metric space $(M^{(i)}, d^{(i)})$, which are updated according to the rules

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} + \lambda_{n+1}^{(i)} \left\{ F^{(i)}\left(\theta_n^{(1)}, \ldots, \theta_n^{(N)}\right) + U_{n+1}^{(i)} \right\}, \tag{4.1}$$

where, for each i, the conditions **B1–B4** of Theorem 23 hold. In addition we assume that

$$\frac{\lambda_n^{(i)}}{\lambda_n^{(j)}} \to 0 \quad \text{as} \quad n \to \infty \quad \text{whenever} \quad i < j.$$

This final assumption is what makes the algorithm multiple-timescale. Write $\theta_n = (\theta_n^{(1)}, \ldots, \theta_n^{(N)})$; in the sequel it will also be convenient to write $\theta_n^{(<i)}$ for the vector

$(\theta^{(1)}, \ldots, \theta^{(i-1)})$.

We follow Borkar (1997) in establishing a different timescale corresponding to each process. For $i, j \in 1, \ldots, N$ let

$$\tau_n^{(j)} = \sum_{k=1}^{n} \lambda_k^{(j)}, \qquad m^{(j)}(t) = \sup\{n \geq 0 : t_n^{(j)} \leq t\},$$

and let $\Theta^{(i,j)}(t)$ be the interpolation of the process $\theta_n^{(i)}$ on the jth timescale, i.e.

$$\Theta^{(i,j)}(\tau_n^{(j)} + s) = \theta_n^{(i)} + \frac{s}{\tau_{n+1}^{(j)} - \tau_n^{(j)}}(\theta_{n+1}^{(i)} - \theta_n^{(i)}) \quad \text{for} \quad 0 \leq s \leq \lambda_{n+1}^{(j)}.$$

We start by considering the Nth timescale, and the interpolations on this timescale $\Theta^{(i,N)}(t)$. Rewrite the stochastic approximation processes (4.1) in the form

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} + \lambda_{n+1}^{(N)} \tilde{U}_{n+1}^{(i,N)} \quad \text{for} \quad i < N,$$
$$\theta_{n+1}^{(N)} = \theta_n^{(N)} + \lambda_{n+1}^{(N)} \left(F^{(N)}(\theta_n) + U_{n+1}^{(N)} \right),$$

where for $i < N$ we have implicitly defined

$$\tilde{U}_{n+1}^{(i,N)} = \frac{\lambda_{n+1}^{(i)}}{\lambda_{n+1}^{(N)}} \left(F^{(i)}(\theta_n) + U_{n+1}^{(i)} \right).$$

For $i < N$, and any n,

$$\sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(N)} \tilde{U}_{l+1}^{(i,N)} \right\| : k = n+1, \ldots, m^{(N)}(\tau_n^{(N)} + T) \right\} \tag{4.2}$$
$$= \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(N)} \frac{\lambda_{l+1}^{(i)}}{\lambda_{l+1}^{(N)}} \left(F^{(i)}(\theta_l) + U_{n+1}^{(i)} \right) \right\| : k = n+1, \ldots, m^{(N)}(\tau_n^{(N)} + T) \right\}$$
$$\leq \sum_{l=n}^{m^{(N)}(\tau_n^{(N)}+T)} \lambda_{l+1}^{(N)} \sup_{k \geq n+1} \left\{ \frac{\lambda_{k+1}^{(i)}}{\lambda_{k+1}^{(N)}} \| F^{(i)}(\theta_k) \| \right\}$$
$$+ \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(i)} U_{l+1}^{(i)} \right\| : k = n+1, \ldots, m^{(N)}(\tau_n^{(N)} + T) \right\}.$$

As $n \to \infty$ the second term converges to zero, by assumption **B4** and the fact that $m^{(N)}(t) \leq m^{(i)}(t)$ for sufficiently large t (since $\tau_n^{(i)} \leq \tau_n^{(N)}$ for sufficiently large n). Also $\lambda_k^{(i)}/\lambda_k^{(N)} \to 0$ while $F^{(i)}(\theta_k)$ is bounded and, from the definitions of $t_n^{(N)}$ and $m^{(N)}$,

$$\sum_{l=n}^{m^{(N)}(t_n^{(N)}+T)} \lambda_{l+1}^{(N)} \approx T.$$

Therefore the limit of the quantity (4.2) as $n \to \infty$ must be zero. Taking $\tilde{U}_n^{(N,N)} = U_n^{(N)}$ we see that

$$\lim_{n\to\infty} \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(N)} \tilde{U}_{l+1}^{(i,N)} \right\| : k = n+1, \ldots, m^{(N)}(\tau_n^{(N)} + T) \right\} = 0$$

for all i. Thus Proposition 17 shows that the interpolations $\Theta^{(\cdot,N)}(t)$ are asymptotic pseudotrajectories of the semiflow induced by the differential equations

$$\begin{aligned} \dot{X}^{(i)} &= 0 \quad \text{for} \quad i < N \\ \dot{X}^{(N)} &= F^{(N)}(X) \end{aligned} \tag{4.3}$$

At this point we need to make the following assumption:

A(N) *There exists a Lipschitz continuous function $\xi^{(N)}(\theta^{(<N)})$ such that, for any initial conditions $(\theta^{(<N)}, \theta^{(N)})$, trajectories of the differential equations (4.3) converge to the point $(\theta^{(<N)}, \xi^{(N)}(\theta^{(<N)}))$.*

It therefore follows from Proposition 21 that the possible limit points of an asymptotic pseudotrajectory to the semiflow induced by (4.3) are the set of all points

$$(\theta^{(<N)}, \xi^{(N)}(\theta^{(<N)})),$$

where $\theta^{(<N)}$ can take any value. In terms of our stochastic approximation processes,

$$\|\theta_n - (\theta_n^{(<N)}, \xi^{(N)}(\theta_n^{(<N)}))\| \to 0 \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

Now consider the timescale corresponding to $\tau_n^{(N-1)}$, and the interpolations $\Theta^{(i,N-1)}(t)$ for $i < N$. Rewrite the stochastic approximation processes (4.1) in the form

$$\begin{aligned} \theta_{n+1}^{(i)} &= \theta_n^{(i)} + \lambda_{n+1}^{(N-1)} \tilde{U}_{n+1}^{(i,N-1)} \quad \text{for} \quad i < N-1 \\ \theta_{n+1}^{(N-1)} &= \theta_n^{(N-1)} + \lambda_{n+1}^{(N-1)} \left\{ F^{(N-1)}(\theta_n^{(<N)}, \xi^{(N)}(\theta_n^{(<N)})) + \tilde{U}_{n+1}^{(N-1,N-1)} \right\} \end{aligned}$$

The implicit definition of $\tilde{U}_{n+1}^{(i,N-1)}$ for $i < N-1$ is equivalent to that of $\tilde{U}_{n+1}^{(i,N)}$, and so we can proceed as before. On the other hand we have implicitly defined

$$\tilde{U}_{n+1}^{(N-1,N-1)} = F^{(N-1)}(\theta_n) - F^{(N-1)}(\theta_n^{(<N)}, \xi^{(N)}(\theta^{(<N)})) + U_{n+1}^{(N-1)}.$$

However, we have already shown that as $n \to \infty$

$$\left\| 0_n - (0_n^{(<N)}, \xi^{(N)}(0_n^{(<N)})) \right\| \to 0,$$

and we have assumed that $F^{(N-1)}$ is Lipschitz continuous, so

$$\left\| F^{(N-1)}(0_n) - F^{(N-1)}(0_n^{(<N)}, \xi^{(N)}(0_n^{(<N)})) \right\| \to 0.$$

Therefore

$$\begin{aligned}
&\sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(N-1)} \tilde{U}_{l+1}^{(N-1,N-1)} \right\| \; : \; k = n+1, \ldots, m^{(N-1)}(\tau_n^{(N-1)} + T) \right\} \\
&\leq \sum_{l=n}^{m^{(N-1)}(\tau_n^{(N-1)}+T)} \lambda_{l+1}^{(N-1)} \sup_{k \geq n+1} \left\| F^{(N-1)}(0_n) - F^{(N-1)}(0_n^{(<N)}, \xi^{(N)}(0^{(<N)})) \right\| \\
&\qquad + \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(N-1)} U_{l+1}^{(N-1)} \right\| \; : \; k = n+1, \ldots, m^{(N-1)}(\tau_n^{(N-1)} + T) \right\},
\end{aligned}$$

and again (since $\sum_{l=n}^{m^{(N-1)}(\tau_n^{(N-1)}+T)} \lambda_{l+1}^{(N-1)} \approx T$) we see that

$$\lim_{n \to \infty} \sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_{l+1}^{(N-1)} \tilde{U}_{l+1}^{(i,N-1)} \right\| \; : \; k = n+1, \ldots, m^{(N-1)}(\tau_n^{(N-1)} + T) \right\} = 0$$

for $i = 1, \ldots, N-1$, and so the interpolations $\Theta^{(<N,N-1)}(t)$ are an asymptotic pseudotrajectory of the flow defined by the differential equations

$$\begin{aligned}
\dot{X}^{(i)} &= 0 \quad \text{for} \quad i < N-1 \\
\dot{X}^{(N-1)} &= F^{(N-1)}\left(X^{(<N)}, \xi^{(N)}(X^{(<N)})\right)
\end{aligned} \tag{4.4}$$

We need to make an assumption analogous to **A(N)** above:

A(N-1) *There exists a Lipschitz continuous function* $\xi^{(N-1)}(0^{(<N-1)})$ *such that, for any initial conditions* $(0^{(<N-1)}, 0^{(N-1)})$, *trajectories of the differential equations (4.4) converge to the point* $(0^{(<N-1)}, \xi^{(N-1)}(0^{(<N-1)}))$.

Defining

$$\Xi^{(\geq N-1)}(0^{(<N-1)}) = \left(\xi^{(N-1)}(0^{(<N-1)}), \xi^{(N)}\left(0^{(<N-1)}, \xi^{(N-1)}(0^{(<N-1)})\right)\right),$$

it follows that

$$\left\|\theta_n - \left(\theta_n^{(<N-1)}, \Xi^{(\geq N-1)}(\theta_n^{(<N-1)})\right)\right\| \to 0 \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

We proceed recursively for each $j \geq 2$, noting that the interpolated processes $\Theta^{(<j+1,j)}$ are asymptotic pseudotrajectories of the semiflow induced by

$$\begin{aligned} \dot{X}^{(i)} &= 0 \quad \text{for} \quad i < j \\ \dot{X}^{(j)} &= F^{(j)}(X^{(<j+1)}, \Xi^{(\geq j+1)}(X^{(<j+1)})) \end{aligned} \tag{4.5}$$

For each $j \geq 2$ we need to make the assumption

A(j) *There exists a Lipschitz continuous function $\xi^{(j)}(\theta^{(<j)})$ such that, for any initial conditions $(\theta^{(<j)}, \theta^{(j)})$, trajectories of the differential equations (4.5) converge to the point $(\theta^{(<j)}, \xi^{(j)}(\theta^{(<j)}))$.*

Then defining

$$\Xi^{(\geq j)}(\theta^{(<j)}) = \left(\xi^{(j)}(\theta^{(<j)}), \Xi^{(\geq j+1)}\left(\theta^{(<j)}, \xi^{(j)}(\theta^{(<j)})\right)\right),$$

it follows that, for $2 \leq j \leq N$,

$$\left\|\theta_n - \left(\theta_n^{(<j)}, \Xi^{(\geq j)}(\theta_n^{(<j)})\right)\right\| \to 0 \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

Finally, it follows that on the slowest timescale the interpolated process $\Theta^{(1,1)}(t)$ is an asymptotic pseudotrajectory of the semiflow induced by

$$\dot{X}^{(1)} = F^{(1)}\left(X^{(1)}, \Xi^{(\geq 2)}(X^{(1)})\right)$$

We have therefore proved the following theorem:

Theorem 40 *Consider a multiple-timescales stochastic approximation process (4.1). If assumptions* **A(2)–A(N)** *hold then, almost surely,*

$$\left\|\theta_n^{(>1)} - \Xi^{(\geq 2)}(\theta_n^{(1)})\right\| \to 0 \quad \textit{as} \quad n \to \infty$$

and a suitable continuous time interpolation of the process $\{\theta_n^{(1)}\}_{n\geq 0}$ is an asymptotic pseudotrajectory of the semiflow induced by the ODE

$$\dot{X} = F^{(1)}(X, \Xi^{(\geq 2)}(X))$$

4.2 A multiple-timescales actor–critic algorithm

Theorem 40 allows us to consider a learning algorithm where the players adapt at different rates. We will modify the actor–critic algorithm (3.5) to assume that all players update their strategies on strictly different timescales, and all of these timescales are slower than the rate at which the Q values are learned. The algorithm is as follows:

Multiple-timescales actor–critic algorithm

Each player i selects an action a_n^i using the strategy π_n^i, then updates π_n^i and Q_n^i according to

$$\begin{aligned} \pi_{n+1}^i &= (1-\mu_{n+1}^i)\pi_n^i + \mu_{n+1}^i \beta^i(Q_n^i) \\ Q_{n+1}^i(a^i) &= Q_n^i(a^i) + \lambda_{n+1}\mathbb{I}_{\{a_n^i=a^i\}}(R_n^i - Q_n^i(a^i)), \quad \text{for } a^i \in A^i, \end{aligned} \tag{4.6}$$

where $\{\lambda_n\}_{n\geq 1}$ and $\{\mu_n^i\}_{n\geq 1}$ are deterministic sequences satisfying the standard conditions (1.15), and $\mu_n^i/\lambda_n \to 0$ as $n \to \infty$. Furthermore, $\mu_n^i/\mu_n^j \to 0$ whenever $i < j$.

As before, R_n^i is the reward obtained by player i at step n, and $\beta^i(Q_n^i)$ is player i's smooth best response given the value estimates Q_n^i. The last condition says that each player is adapting their strategy on a different timescale (although all players still learn the Q values at the same fast timescale).

The first thing to note about this algorithm is that the same argument as for the simple actor–critic algorithm (3.5) will suffice to show the following.

Theorem 41 *If the multiple-timescales actor–critic algorithm (4.6) converges to a fixed point*

$$(Q_n, \pi_n) \to (\tilde{Q}, \tilde{\pi}) \quad \text{as} \quad n \to \infty$$

then $\tilde{Q}^i(a^i) = r^i(a^i, \tilde{\pi}^{-i})$ *and* $\tilde{\pi}$ *is a Nash distribution.*

However to use Theorem 40 we need to check that assumptions A(2)–A(N)

are satisfied. We start by noting that the ODE for the Q values,

$$\tfrac{\mathrm{d}}{\mathrm{d}t}Q^i(a^i) = \pi^i(a^i)\left\{r^i(a^i, \pi^{-i}) - Q^i(a^i)\right\} \quad \text{for all } i, a^i,$$

has a unique globally attracting fixed point $Q = \underline{r}(\pi)$ (recall definition (3.6)). The remainder of our analysis will assume that the Q values are accurate, and so smooth best responses are taken with respect to opponent strategies (as opposed to Q values). It is also clear that the ODE

$$\dot{\pi}^N = \beta^N(\pi^1, \ldots, \pi^{N-1}) - \pi^N,$$

for fixed $\pi^{<N} = (\pi^1, \ldots, \pi^{N-1})$, has a globally attracting point, $\beta^N(\pi^{<N})$, so the assumptions **A(2)–A(N)** may fail only for intermediate players that are not the fastest or slowest (no assumption need be made about the slowest timescale).

Assumption C *For each $i = 2, \ldots, N-1$ there exists a Lipschitz function b^i such that $b^i(\pi^1, \ldots, \pi^{i-1})$ is the globally asymptotically stable equilibrium point of the ODE*

$$\dot{\pi}^i = \beta^i\left(\pi^{<i}, B^{>i}(\pi^{\leq i})\right) - \pi^i$$

where we recursively define

$$B^{>(N-1)}(\pi^{\leq(N-1)}) = \beta^N(\pi^{\leq(N-1)})$$

$$B^{>i}(\pi^{\leq i}) = \left[b^{i+1}(\pi^{\leq i}), B^{>(i+1)}\left(\pi^{\leq i}, b^{i+1}(\pi^{\leq i})\right)\right]$$

Effectively this will tell us that, for any i, if we fix the strategies for players $1, \ldots, i$ then almost surely

$$\pi_n^{>i} \to B^{>i}(\pi^{\leq i}).$$

This convergence assumption is fairly restrictive, although it does not prevent the application of this algorithm to several different games (see Sections 4.3–4.4 below). It allows us to use Theorem 40 to characterise the asymptotic behaviour of the algorithm (4.6).

Theorem 42 *Under Assumption C, the strategies π_n played during the multiple-timescales actor-critic algorithm (4.6) satisfy*

$$\left\|(\pi_n^2, \ldots, \pi_n^N) - B^{>1}(\pi_n^1)\right\| \to 0 \quad as \quad n \to \infty \quad a.s.$$

and a suitable continuous time interpolation of the π_n^1 is an asymptotic pseudotrajectory of the semiflow induced by the ODE

$$\dot{\pi}^1 = \beta^1\left(B^{>1}(\pi^1)\right) - \pi^1$$

PROOF Since the $Q_n^i(a^i) \to r^i(a^i, \pi^{-i})$ whenever π is fixed, the proof is immediate from Theorem 40 and Assumption C.

This result means that to analyse the multiple-timescales algorithm in a particular game, or class of games, it suffices to show that Assumption C is satisfied and to analyse the behaviour of the slowest player under the assumption that all other players play the strategy dictated by the function $B^{>1}$.

REMARK We can consider this system as relating to a multiple-timescales singular perturbation of the smooth best response dynamics:

$$\begin{aligned}
\dot{\pi}^1 &= \epsilon^{(1)}\beta^1(\pi^{-1}) - \pi^1, \\
\dot{\pi}^2 &= \epsilon^{(2)}\left(\beta^2(\pi^{-2}) - \pi^2\right), \\
&\vdots \\
\dot{\pi}^N &= \left(\beta^N(\pi^{-N}) - \pi^N\right),
\end{aligned}$$

with $\epsilon^{(i)} = o(\epsilon^{(i+1)})$ as $\epsilon^{(i)} \to 0$. Consideration of this system may indicate how to relax Assumption C.

REMARK It is a trivial extension to this algorithm to consider a setting where the players also learn the Q values at different rates to each other. If we assume that

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \lambda_{n+1}^i \mathbb{I}_{\{a_n^i = a^i\}}(R_n^i - Q_n^i(a^i)),$$

then all that is required is for $\mu_n^i/\lambda_n^i \to 0$ as $n \to \infty$. This is because the values Q^i only have any direct bearing on the strategy π^i of player i. The condition $\mu_n^i/\lambda_n^i \to 0$ corresponds to the players being 'cautious', in that they will adapt their strategy more slowly than they will adjust assessments of the values of actions.

4.3 Two-player games

It is easy to see that for 2-player games Assumption C is vacuous, since there are no intermediate players (each player is either the fastest or the slowest). Thus it is sufficient to analyse the ODE

$$\dot{\pi}^1 = \beta^1(\beta^2(\pi^1)) - \pi^1 \tag{4.7}$$

We have a positive convergence theorem for two major classes of 2-player games: zero-sum games and partnership games.

Proposition 43 *For both 2-player zero-sum games and 2-player partnership games the ODE (4.7) admits a Lyapunov function for the set of Nash distributions.*

PROOF For zero-sum games consider the function $U(\pi^1) = r^1(\pi^1, \beta^2(\pi^1)) + \tau v^1(\pi^1) - \tau v^2(\beta^2(\pi^1))$. We see that

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}U(\pi^1) &= \left\{r^1(\cdot, \beta^2(\pi^1)) + \tau \nabla v^1(\pi^1)\right\} \cdot \dot{\pi}^1 + \left\{r^1(\pi^1, \cdot) - \tau \nabla v^2(\beta^2(\pi^1))\right\} \cdot \tfrac{\mathrm{d}}{\mathrm{d}t}\beta^2(\pi^1) \\ &= \tau \left\{-\nabla v^1(\beta^1(\beta^2(\pi^1))) + \nabla v^1(\pi^1)\right\} \cdot \left\{\beta^1(\beta^2(\pi^1)) - \pi^1\right\} \\ &\quad - \left\{r^2(\pi^1, \cdot) + \nabla v^2(\beta^2(\pi^1))\right\} \cdot \tfrac{\mathrm{d}}{\mathrm{d}t}\beta^2(\pi^1).\end{aligned}$$

By the definitions of the smooth best responses, the second term is 0 and $\frac{\mathrm{d}}{\mathrm{d}t}U(\pi^1) \geq 0$ with equality only at Nash distributions. So $-U(\pi^1)$ is a Lyapunov function for the set of Nash distributions.

For partnership games consider the function $V(\pi^1) = r(\pi^1, \beta^2(\pi^1)) + \tau v^1(\pi^1) +$

$\tau v^2(\beta^2(\pi^1))$. Again

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V(\pi^1) &= \left\{r(\cdot,\beta^2(\pi^1)) + \tau\nabla v^1(\pi^1)\right\}\cdot\dot{\pi}^1 + \left\{r(\pi^1,\cdot) + \tau\nabla v^2(\beta^2(\pi^1))\right\}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\beta^2(\pi^1)\\ &= \tau\left\{-\nabla v^1(\beta^1(\beta^2(\pi^1))) + \nabla v^1(\pi^1)\right\}\cdot\left\{\beta^1(\beta^2(\pi^1)) - \pi^1\right\}\\ &\geq 0\end{aligned}$$

and so $-V(\pi^1)$ is a Lyapunov function for the set of Nash distributions.

Corollary 44 *Under the multiple-timescales actor–critic algorithm (4.6) in a 2-player zero-sum game, the strategies of the players will converge to the unique Nash distribution. In an N-player partnership game with finite or countably many Nash distributions, the strategies will converge to one of these Nash distributions.*

PROOF The proof of this corollary is immediate from Theorem 42, Proposition 43, and Proposition 21.

Thus we have asymptotic convergence results which are comparable to those for smooth fictitious play, and for our simple actor–critic algorithm (3.5). However a proof of convergence for general N-player partnership games is not available, since in this framework it is likely that for a fixed strategy of the slow players there will be several equilibria to which the fast players may converge, and Assumption C will not be satisfied.

4.4 Some difficult games

We now consider the multiple-timescales actor–critic algorithm applied in the two classic examples of difficult games for learning algorithms: the Shapley game (3.2) and the 3-player matching pennies game (3.3). We start by proving convergence of our algorithm in a generalisation of the latter game, then show convergence of our algorithm for the Shapley game.

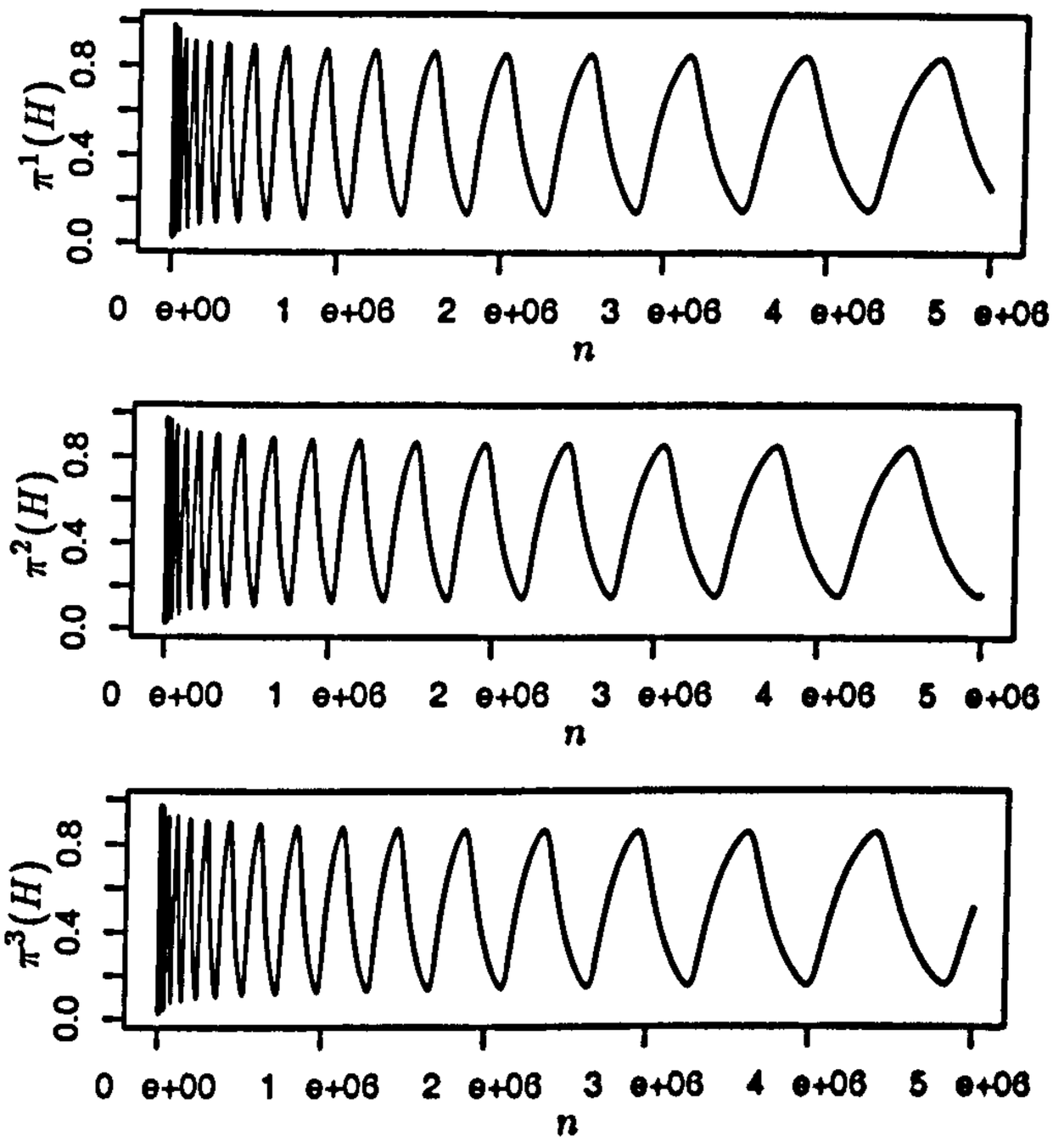


Figure 4.1: Non-convergent strategies in the 3-player matching pennies game, over 5×10^5 iterations of the single-timescale actor–critic algorithm (3.5), using Boltzmann smoothing ($\tau = 0.1$), with $\lambda_n = (n + 100)^{-0.55}$ and $\mu_n = (n + 100)^{-0.8}$.

4.4.1 N-player matching pennies

Our generalisation of Jordan's game (Jordan 1993) is the N-player matching pennies game, in which each player can choose to play 'heads' (H) or 'tails' (T) and the reward to player i depends only on the actions a^i and a^{i+1}, where $i + 1$ is calculated modulo N. The reward structure is

$$\begin{aligned} r^i(\underline{a}) &= \mathbb{I}_{\{a^i = a^{i+1}\}} \quad \text{for} \quad i = 1, \ldots, N-1, \\ r^N(\underline{a}) &= \mathbb{I}_{\{a^N \neq a^1\}}. \end{aligned} \tag{4.8}$$

The cyclical nature of this game allows the easy verification of Assumption C. As long as player 1's strategy is fixed then player N's strategy will converge to $\beta^N(\pi^{-N})$ since this only depends on π^1. Similarly, under the assumption that player one is fixed and player N has calibrated, it is clear that player $(N-1)$'s

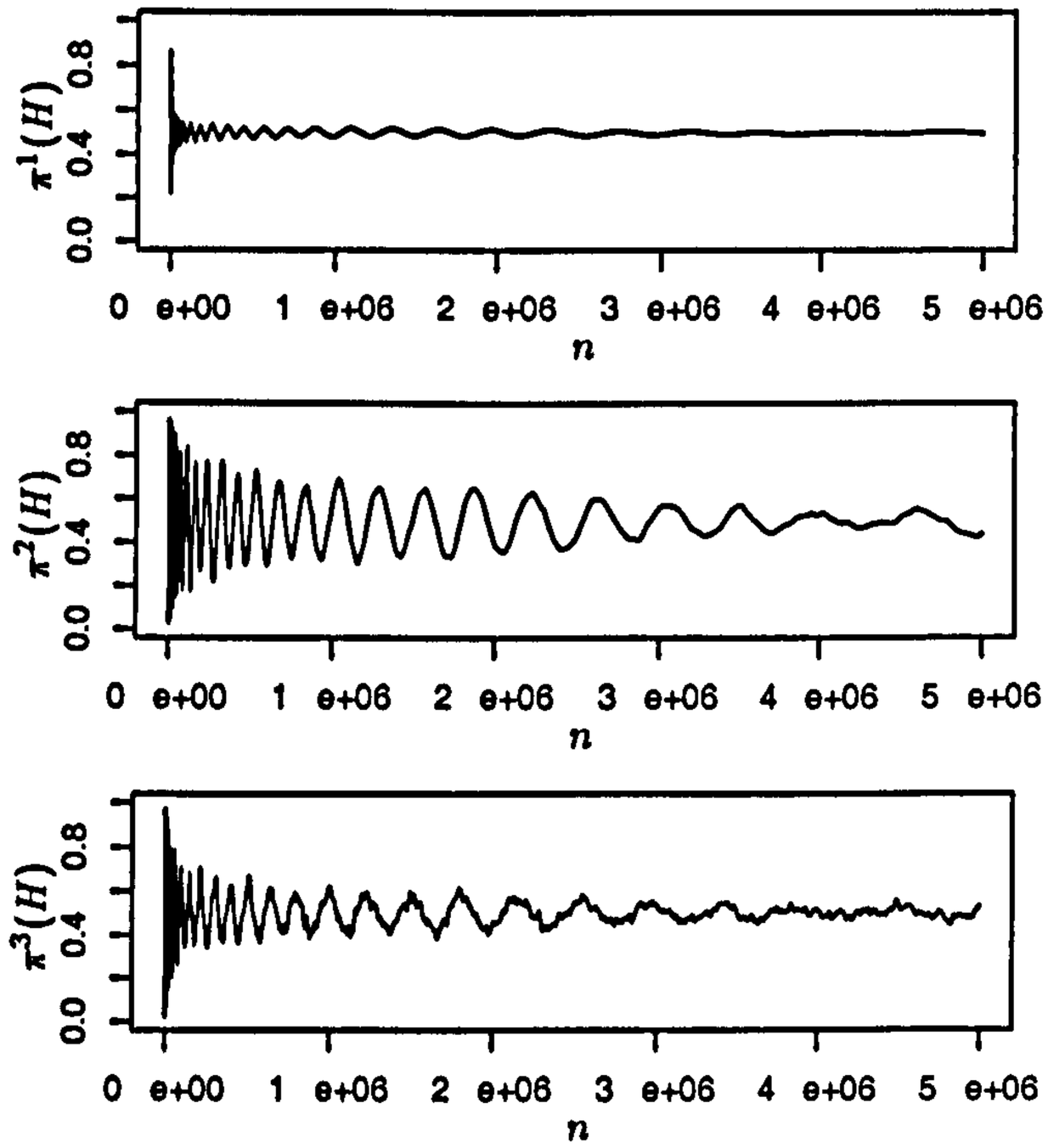


Figure 4.2: Convergent strategies in the 3-player matching pennies game, over 5×10^5 iterations of the multiple-timescales actor–critic algorithm (4.6), using Boltzmann smoothing ($\tau = 0.1$), with $\lambda_n = (n + 100)^{-0.55}$, $\mu_n^1 = (n + 100)^{-1}$, $\mu_n^2 = (n + 100)^{-0.8}$ and $\mu_n^3 = (n + 100)^{-0.6}$.

strategy will converge to $\beta^{N-1}(\pi^{-(N-1)})$, since this depends only on $\pi^N = \beta^N(\pi^{-N})$ which is fixed. This is repeated, so that whenever player 1's strategy is fixed the strategies of the faster players must converge to the unique best responses. By Theorem 42 it suffices to consider the ODE

$$\dot{\pi}^1 = \beta^1(\beta^2(\ldots(\beta^N(\pi^1))\ldots)) - \pi^1.$$

Assume that the smooth best responses are monotonic in the payoffs i.e. $r^i(a^i) > r^i(b^i) \Rightarrow \beta^i(r^i)(a^i) > \beta^i(r^i)(b^i)$ (a sufficient condition for this to be the case is for each smoothing function v^i to be invariant under permutations of the actions). Thus if $\pi^1(II) > 1/2$ we must have $\beta^N(\pi^1)(II) < 1/2$ and so, in turn,

$$\beta^i(\beta^{i+1}(\ldots(\beta^N(\pi^1))\ldots))(II) < 1/2$$

for each $i = 1, \ldots, N$. So for $\pi^1(II) > 1/2$ it is the case that $\dot{\pi}^1(II) < 0$. Similarly if $\pi^1(H) < 1/2$ then $\dot{\pi}^1(II) > 0$, and so it follows that the Nash distribution $\pi^i(H) = 1/2$ is a globally attracting fixed point. Therefore the strategies under the multiple-timescales actor–critic algorithm will converge to this Nash distribution.

REMARK We have shown that the multiple-timescales algorithm (4.6) will converge almost surely to the Nash distribution of the matching pennies game provided that learning rates of the players are ordered in the same way as the players are ordered in the game. In fact it is not difficult to see that this specific ordering is unnecessary, and any ordering of the players will suffice; see Section 4.5.

We conclude our discussion of the N-player matching pennies game with sample learning trajectories, shown in Figs. 4.1 and 4.2. As predicted by Theorem 34, in conjunction with the results of Benaïm and Hirsch (1999) on the smooth best response dynamics in this game, the single-timescale algorithm (3.5) cycles persistently (Fig. 4.1). However, the multiple-timescales algorithm (4.6) is converging slowly towards the Nash distribution where $\pi^i(II) = 1/2$ for each i (Fig. 4.2). This is in agreement with the theoretical results of this section. Also visible in Fig. 4.2 is the different behaviour of the three players. Player 1, with learning parameters that decrease quickly towards 0, does not adjust π^1 very rapidly. However, player 3, with slowly decreasing learning parameters, displays behaviour with a fairly persistent level of stochasticity, although π^3 is always close to being a smooth best response to π^1 (not evident from the plot).

4.4.2 Shapley's game

Shapley's game (3.2) is a 2-player game, where each player has three actions. A player gets a point if their opponent plays an action 1 greater (modulo 3) and gets no point otherwise. Without loss of generality (due to the symmetry of the game) we assume that player 1 is the slower, and since it is a 2-player game Assumption

C is irrelevant (as observed previously). So we simply need to analyse the ODE

$$\dot{\pi}^1 = \beta^1(\beta^2(\pi^1)) - \pi^1. \tag{4.9}$$

Note $\pi^1(3) = 1 - \pi^1(1) - \pi^1(2)$, so that this defines a planar semiflow. Therefore if the divergence of the semiflow in $(\pi^1(1), \pi^1(2))$-space is negative then the solutions of the ODE must converge to a fixed point.

For simplicity we assume that players use the Boltzmann distribution (1.16) for their smooth best responses. It follows that

$$\beta^i(\pi^{-i})(a) = \frac{e^{\pi^{-i}(a+1)/\tau}}{\sum_{a' \in A} e^{\pi^{-i}(a')/\tau}}.$$

Define $\rho(a) = (\pi^1(a) - \pi^1(3))/\tau$ for $a = 1, 2$, and let

$$\pi^2 = \beta^2(\pi^1) = \frac{1}{1 + e^{\rho(1)} + e^{\rho(2)}}(e^{\rho(2)}, 1, e^{\rho(1)}). \tag{4.10}$$

By the chain rule applied to (4.9),

$$\text{Div} = \sum_{a=1}^{2} \frac{\partial \dot{\pi}^1(a)}{\partial \pi^1(a)} = \sum_{a=1}^{2} \sum_{a'=1}^{3} \sum_{b=1}^{2} \frac{\partial \beta^1(\pi^2)(a)}{\partial \pi^2(a')} \frac{\partial \pi^2(a')}{\partial \rho(b)} \frac{\partial \rho(b)}{\partial \pi^1(a)} - 2,$$

so to calculate the value of this sum we first calculate the component partial derivatives:

$$\frac{\partial \beta^1(\pi^2)(a)}{\partial \pi^2(a')} = \frac{e^{\pi^2(a')/\tau}\left(\mathbb{I}_{\{a'=a+1\}} \sum_{b' \in A} e^{\pi^2(b')/\tau} - 1\right)}{\tau\left(\sum_{b' \in A} e^{\pi^2(b')/\tau}\right)^2},$$

$$\frac{\partial \pi^2}{\partial \rho(1)} = \frac{e^{\rho(1)}}{(1 + e^{\rho(1)} + e^{\rho(2)})^2}\,(-e^{\rho(2)}, \quad -1, \; 1 + e^{\rho(2)}),$$

$$\frac{\partial \pi^2}{\partial \rho(2)} = \frac{e^{\rho(2)}}{(1 + e^{\rho(1)} + e^{\rho(2)})^2}\,(1 + e^{\rho(1)}, \; -1, \quad -e^{\rho(1)}),$$

$$\frac{\partial \rho(b)}{\partial \pi^1(a)} = (1 + \mathbb{I}_{\{a=b\}})/\tau,$$

where the last derives from the fact that $\pi^1(3) = 1 - \pi^1(1) - \pi^1(2)$ and so

$$\rho(1) = (2\pi^1(1) + \pi^1(2) - 1)/\tau, \qquad \rho(2) = (\pi^1(1) + 2\pi^1(2) - 1)/\tau.$$

Substituting all of these into the expression for the divergence, we get that

$$\tau^2 \left(\sum_{a=1}^{3} e^{\pi^2(a)/\tau} \right)^2 \left(1 + e^{\rho(1)} + e^{\rho(2)}\right)^2 \times (\mathrm{Div} + 2)$$

$$= e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} (e^{\rho(1)} e^{\rho(2)} - 2e^{\rho(1)} - 2e^{\rho(2)})$$

$$+ e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} (e^{\rho(2)} - 2e^{\rho(1)} - 2e^{\rho(1)} e^{\rho(2)})$$

$$+ e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} (e^{\rho(1)} - 2e^{\rho(2)} - 2e^{\rho(1)} e^{\rho(2)})$$

Recalling the expression (4.10) for π^2, this shows that

$$\tau^2 \left(\sum_{a=1}^{3} e^{\pi^2(a)/\tau} \right)^2 \times (\mathrm{Div} + 2)$$

$$= e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} \left\{ \pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) \right\} \tag{4.11}$$

$$+ e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(1)\pi^2(2) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(3) \right\}$$

$$+ e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) \right\}$$

This expression is invariant under a cyclical permutation of actions, so without loss of generality we can assume $\pi^2(1) \leq \pi^2(3)$ and $\pi^2(2) \leq \pi^2(3)$. Initially we assume further that $\pi^2(1) \leq \pi^2(2) \leq \pi^2(3)$, so that

$$\pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) < 0,$$

$$\pi^2(1)\pi^2(2) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(3) < 0.$$

If $\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) < 0$ we are done. Otherwise

$$e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) \right\}$$

$$\leq e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) \right\},$$

and the expression in (4.11) is bounded above by

$$e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} \left\{ \pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) \right\}$$

$$+ e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \left\{ -\pi^2(1)\pi^2(2) - \pi^2(2)\pi^2(3) - 4\pi^2(1)\pi^2(3) \right\},$$

which is clearly negative. A similar argument works with the assumption $\pi^2(2) \leq \pi^2(1) \leq \pi^2(3)$, and so the expression in (4.11) is always negative. This shows that

$$\mathrm{Div} = \sum_{a=1}^{2} \frac{\partial \dot{\pi}^1(a)}{\partial \pi^1(a)} \leq -2.$$

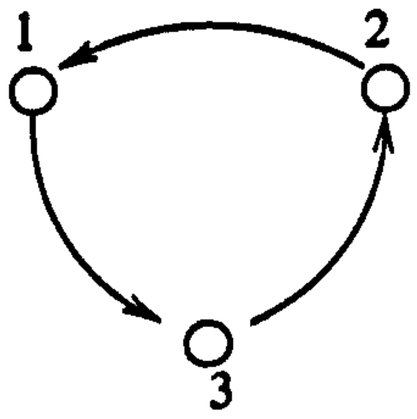
Since we have a planar semiflow with negative divergence the system must converge to a fixed point; there is a unique fixed point, at the Nash distribution (Cowan 1992), so this point must be globally attracting. Therefore from Theorem 42 it follows that the learning algorithm (4.6) will converge with probability 1 to the Nash distribution of the Shapley game.

4.5 A graphical analysis

It is of interest to know when the convergence assumption (Assumption C) will hold. One method that can be used to analyse games is to extend the concept of graphical games (Littman *et al.* 2001; Koller and Milch 2003).

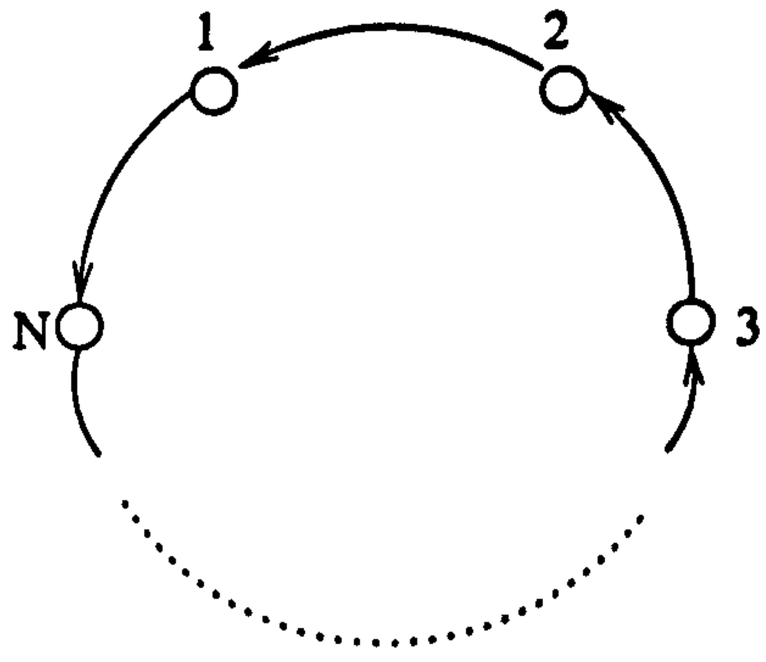
Given a game, we draw a graph with a node corresponding to each player. An arc is drawn from node i to node j if the actions of player i directly affect the rewards received by player j. Thus a simple non-degenerate 2-player game is represented by the graph

and Jordan's 3-player matching pennies game is represented by the graph

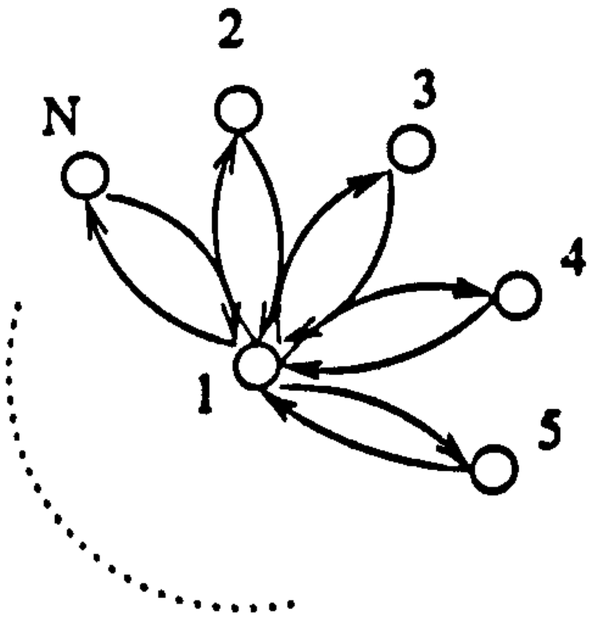


When considering the effects of using multiple timescales, a graph can be drawn. The node representing the slowest player is then removed, along with all arcs connected to it. The remaining graph represents the game as if the strategy of the removed player is fixed. In particular, if the resultant graph is acyclic then Assumption C holds. An obvious class of games for which this holds is games for which the initial graph is acyclic, i.e. a directed tree. A slightly more complicated

class of games where this happens is the class of cyclic games, such as N-player matching pennies (Section 4.4.1), where the graph is a simple directed cycle:



Removing any node leaves a simple directed line of nodes, and convergence clearly follows. A third class is that of star games, where each agent only interacts directly with one central controller:

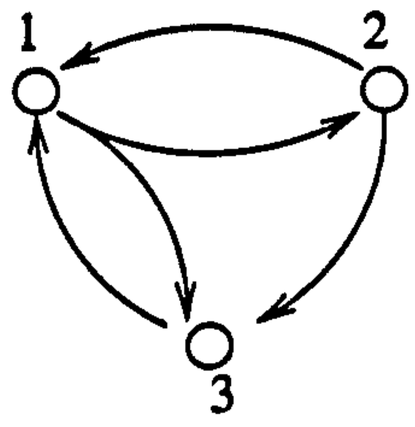


In this case, removing the central player means that the graph is disconnected, and therefore acyclic. This class of games could have applications in areas such as auctions and computer communications.

However, it may be necessary to consider the resulting game more thoroughly. Consider a 3-player game, consisting of a matching pennies game with an apprentice for one of the players:

Player 3's choice:		HEAD		TAIL	
Player 2's choice:		HEAD	TAIL	HEAD	TAIL
Player 1's choice	H	(1,0,1)	(0,1,0)	(0,0,0)	(0,1,0)
	T	(0,1,0)	(0,0,0)	(0,1,0)	(1,0,1)

Players 1 and 2 are playing a matching pennies game, while Player 3 is an apprentice to Player 1: Players 1 and 3 will not get any points unless they both win the matching pennies game (by matching Player 2), while Player 2 scores a point by simply playing the opposite action to Player 1. The actions of Player 3 only directly affect the reward to Player 1, and so the graph of this game is given by



Removal of node 1 leaves a very simple graph, and so Assumption C would hold if Player 1 is slowest. On the other hand, suppose Player 2 is the slowest, resulting in a graph as for a generic 2-player game. For any fixed strategy π^2, Players 1 and 3 face a coordination game similar to that studied in Section 2.3; there is not a unique Nash distribution to such a game, and so Assumption C fails to hold. Finally, however, suppose that Player 3 is slowest. Again a generic 2-player graph arises, but now for fixed π^3 this is a game with a unique Nash distribution, and so Assumption C does hold in this case.

So we see that graphical methods can be useful to help analyse whether the learning algorithm will satisfy Assumption C, and may even help a system engineer to decide on fast and slow learners (in the case of a computer communications for example). However, they are not a complete solution to the problem of determining whether Assumption C will hold, as demonstrated by our example of matching pennies with an apprentice.

4.6 Conclusion

We have extended the actor–critic algorithm of the previous chapter to the case of players that learn at different rates, thus breaking the symmetry that causes

strategies to cycle. This necessitated an extension Borkar's (1997) two-timescales stochastic approximation result to the case of multiple timescales. However, Borkar's simple assumption that the ODE associated with the fast timescale has a unique globally attracting fixed point becomes rather more cumbersome in the multiple-timescales case.

We show that this extended algorithm will still converge for the simple 2-player games in which the simpler actor–critic algorithm (3.5) (and stochastic fictitious play) will converge, and also that the algorithm converges to the unique Nash distribution for Shapley's game and for a generalisation of Jordan's 3-player pennies game. This is the first (sensible) algorithm for which this is known to happen.

A further generalisation to the case of 'cautious' players, where each player uses different learning parameter schedules for the Q values, gives identical asymptotical results. The only constraint in this case is that no two players learn at identical rates, and that all players adapt their strategies at a slower rate than they learn their Q values. Thus the players in this algorithm are truly independent: if players use a continuous distribution to choose random $\rho^i_\lambda, \rho^i_\mu \in (0.5, 1]$ with $\rho^i_\lambda < \rho^i_\mu$ and set $\lambda^i_n = (n + C_\lambda)^{-\rho^i_\lambda}$, $\mu^i_n = (n + C_\mu)^{-\rho^i_\mu}$ for constant $C_\lambda, C_\mu > 0$, then the appropriate conditions will be satisfied without any coordination between players at all.

Analysing when the convergence conditions of the algorithm will hold can prove difficult, but a graphical analysis may help in certain situations. Four classes of games for which Assumption C holds are 2-player games, games for which the graph is a tree, cyclical games, and star games (if the 'hub' is the slowest player).

Chapter 5

Q-learning in normal form games

In this chapter we study 'individual Q-learning', where players do not maintain an explicit strategy, and simply utilise a standard value-based reinforcement learning algorithm. In Section 5.2 we will show that individual Q-learners will converge to Nash distribution in a 2-player zero-sum game. We consider a multiple-timescales Q-learning algorithm in Section 5.3, proving that strategies converge to Nash distribution in several classes of game.

Ideas from this chapter were presented in the Multi-Agent Learning workshop at NIPS 2002.

5.1 Individual Q-learning

The model we consider is a simple one, under which players simply play a distribution based upon their current Q values, and adjust these Q values towards the reward observed. While the standard method of adjusting Q-values, as used in the actor-critic algorithm (3.5), would be an obvious choice, Fudenberg and Levine (1998) suggest that the update to $Q_n^i(a^i)$ should be divided by the probability of playing a^i (see (5.1)). This is motivated partly by the fact that if an action is played with low probability, it will not be played often, and any observed changes in value need to have a significant effect. Further motivation is that for

2-player games this results in trajectories that are closely related to the smooth best response dynamics (1.8), and so results can be inferred from these well-studied dynamics. The algorithm to be studied is therefore:

Individual Q-learning algorithm

Each player i selects an action a_n^i using the strategy $\beta^i(Q_n^i)$, then updates Q_n^i according to

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \frac{\lambda_n \mathbb{I}_{\{a_n^i = a^i\}}}{\beta^i(Q_n^i)(a^i)} \left(R_n^i - Q_n^i(a^i)\right), \quad \text{for } a^i \in A^i, \qquad (5.1)$$

where $\{\lambda_n\}_{n\geq 1}$ is a deterministic sequence satisfying the standard conditions (1.15).

Again, the only information used by player i is the action she played and the reward she was given—she sees this reward to be simply a random variable (which is of course dependent on the action played by the opponents).

Proposition 45 *If the individual Q-learning algorithm (5.1) converges to a fixed point*

$$Q_n \to \tilde{Q} \quad as \quad n \to \infty$$

then the strategies $\beta^i(\tilde{Q}^i)$ are a Nash distribution.

PROOF As before, convergence can only occur when the expected change in Q_n is zero. Writing $\beta^{-i}(Q^{-i})$ for the opponent strategies arising from the values Q^{-i}, it is clear that

$$\mathbb{E}\left[\{Q_{n+1}^i(a^i) - Q_n^i(a^i)\}/\lambda_n \,|\, Q_n = Q\right] = r^i(a^i, \beta^{-i}(Q^{-i})) - Q^i(a^i),$$

and so we must have $\tilde{Q}^i = r^i(\cdot, \beta^{-i}(\tilde{Q}^{-i}))$. Therefore, for each i, $\beta^i(\tilde{Q}^i)$ is a smooth best response to the opponent strategies, and therefore the $\beta^i(\tilde{Q}^i)$ are a Nash distribution.

Following identical arguments to those previously given, it is clear that a continuous time interpolation of the Q values is an asymptotic pseudotrajectory of the semiflow induced by

$$\dot{Q}^i(a^i) = r^i(a^i, \beta^{-i}(Q^{-i})) - Q^i(a^i), \tag{5.2}$$

so long as the Q values remain bounded. Note that division by $\beta^i(Q^i_n)(a^i_n)$ may cause difficulties with boundedness, but if we set $\lambda_n = \min\{\min_i \beta^i(Q^i_n)(a^i_n), \lambda'_n\}$, with $\{\lambda'_n\}_{n\geq 1}$ satisfying (1.15), then there are no difficulties (eventually the boundedness of Q values means that $\lambda_n = \lambda'_n$). An alternative approach would be to use the method of random truncations (Chen and Zhu 1986). However, in practice (Section 5.4) there are no difficulties with boundedness, and so we are content with leaving boundedness as an assumption.

5.2 2-player zero-sum games

Fudenberg and Levine (1998) observe that, for 2-player games, if π evolves according to the smooth best response dynamics (1.8), then $r^i(a^i, \pi^{-i})$ follows trajectories of the Q-learning ODE (5.2). Reversing this argument, if initial Q values are *belief-based*, i.e. satisfy $Q^i(a^i) = r^i(a^i, \pi^{-i})$ for some π, and evolve according to (5.2), then they will behave as if they result from evolution of strategies according to the smooth best response dynamics.

Lemma 46 *In 2-player zero-sum games, trajectories of the ODE (5.2) with belief-based initial conditions will converge to a unique fixed point, corresponding to the unique Nash distribution.*

PROOF Suppose the initial conditions correspond to the beliefs π. The trajectory of the smooth best response dynamics (1.8) starting at π will converge to the unique Nash distribution $\tilde{\pi}$ (Theorem 32), and so the trajectories of $r^i(\pi)$ will converge to $r^i(\tilde{\pi})$. But these trajectories are the trajectories of the Q values, and the result follows.

Despite this lemma, it is not immediate that the updates (5.1) will result in convergence to Nash distribution, since we have only shown that trajectories of (5.2) converge to Nash distribution values for a limited set of initial conditions; we need to analyse the global convergence properties of (5.2). In order to do this, we will prove a general dynamical systems result in order to show that the Q values are asymptotically belief-based. We need the following version of Gronwall's inequality:

Theorem 47 (Gronwall's inequality) *Let X be a Banach space, and $U \subset X$ be an open set. Let $f, g : U \to X$ be continuous functions, and let $y, z : [t, t+T] \to U$ satisfy*

$$\dot{y} = f(y), \quad \dot{z} = g(z).$$

Assume that f is Lipschitz, with constant L, and that there exists a continuous function $\eta : [0, T] \to [0, \infty)$ such that $\|f(z(t+h)) - g(z(t+h))\| \leq \eta(h)$. Then for $h \in [0, T]$

$$\|y(t+h) - z(t+h)\| \leq e^{Lh}\|y(t) - z(t)\| + e^{Lh}\int_0^h e^{-L\tau}\eta(\tau)\,\mathrm{d}\tau.$$

PROOF For $s \in [t, t+T]$,

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}s}\|y(s) - z(s)\| &\leq \|\dot{y}(s) - \dot{z}(s)\| \\
&= \|f(y(s)) - g(z(s))\| \\
&\leq \|f(y(s)) - f(z(s))\| + \|(f(z(s)) - g(z(s))\| \\
&=\leq L\|y(s) - z(s)\| + \eta(s-t).
\end{aligned}$$

Thus for $\tau \in [0, T]$

$$\tfrac{\mathrm{d}}{\mathrm{d}\tau}\left(\mathrm{e}^{-L\tau}\|y(t+\tau) - z(t+\tau)\|\right) \leq \mathrm{e}^{-L\tau}\eta(\tau)$$

and the result follows.

Recalling the definition (3.6), and examining the Q-learning ODE (5.2), we see that the Q values are always attracted towards points $\underline{r}(\beta(Q))$, which lie in the

interior of some lower-dimensional space—the set of all points $\underline{r}(\pi)$ is a compact, convex subset of an affine subspace of the set of possible Q values, and since smooth best responses are used the rewards lie in the interior of the set of values $\underline{r}(\pi)$. We will consider an abstract representation of this system.

Lemma 48 *Consider the ODE*

$$\dot{y} = F(y) - y, \quad y \in \mathbb{R}^m, \tag{5.3}$$

where $F : \mathbb{R}^m \to \mathbb{R}^m$ is a Lipschitz continuous function. Take an affine subspace $S \subset \mathbb{R}^m$ and a compact, convex subset $D \subset S$. Suppose $F(y) \in int_S(D)$ for all $y \in \mathbb{R}^m$, where int_S denotes the interior with respect to the linear subspace S, and further that all trajectories with initial conditions $y_0 \in D$ converge to a unique fixed point y^*. Then y^* is a globally attracting fixed point of the ODE.

PROOF Without loss of generality we can assume $S = \{y \in \mathbb{R}^m : y_{m'+1} = \cdots = y_m = 0\}$ for some $1 \leq m' \leq m$, and that $D = \{y \in S : |y| \leq 1\}$ where $|\cdot|$ denotes Euclidean distance. Thus (5.3) becomes

$$\begin{pmatrix} \dot{y}_1 \\ \vdots \\ \dot{y}'_m \\ \dot{y}_{m'+1} \\ \vdots \\ \dot{y}_m \end{pmatrix} = \begin{pmatrix} F_1(y) \\ \vdots \\ F_{m'}(y) \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y'_m \\ y_{m'+1} \\ \vdots \\ y_m \end{pmatrix}$$

Consider trajectories restricted to the invariant subspace S, and consider $W = |y|^2/2$; it is clear that

$$\dot{W} = y \cdot \dot{y} = y \cdot (F(y) - y) \leq |y||F(y)| - |y|^2.$$

Thus for $|y| \geq 1$ we have $\dot{W} < 0$, since $|F(y)| < 1$, and so there exists an $\epsilon > 0$ such that if $|y| > 1 - \epsilon$ then $\dot{W} < 0$.

Suppose $|y| > 1$ for all time, and so $\dot{W} < 0$ for all time; this implies that W tends to a limit and $\dot{W} \to 0$. Therefore y converges to the set $\{y : |y| \leq 1 - \epsilon\}$, contradicting our assumption that $|y| > 1$ for all time. But then it follows that in finite time we must have $|y| \leq 1$, i.e. $y \in D$. Therefore the trajectory must converge to y^*, since all trajectories starting in D converge to y^*, and so y^* is a globally attracting fixed point of the semiflow restricted to the affine subspace S.

Now consider a general trajectory, not restricted to lie in S. It is clear that the components $y_{m'+1}, \ldots, y_m$ will tend to zero exponentially, so we compare (5.3) to the system that behaves as if these components are already zero. Define $\tilde{F}(y_1, \ldots, y_m) = F(y_1, \ldots, y_{m'}, 0, \ldots, 0)$, and consider the semiflow φ defined by

$$\dot{z} = \tilde{F}(z) - z. \tag{5.4}$$

It is clear, due to the decoupling of $(z_1, \ldots, z_{m'})$ and $(z_{m'+1}, \ldots, z_m)$, that $z_i(t) = z_i(0)e^{-t}$ for $i > m'$, and that y^* is the unique globally attracting fixed point of this semiflow.

Define $f(y) = F(y) - y$ and $g(z) = \tilde{F}(z) - z$. Since F is Lipschitz continuous, with constant C_1 say, it is clear that both f and g are Lipschitz continuous (with constant $C_2 = C_1 + 1$), and also that

$$\|f(z(t+h)) - g(z(t+h))\| \leq C_1 \sum_{j=m'+1}^{m} |z_j(t+h)| = C_3 e^{-(t+h)}.$$

Thus by Gronwall's inequality, if y solves (5.3) and z solves (5.4),

$$\|y(t+h) - z(t+h)\| \leq e^{C_2 h}\|y(t) - z(t)\| + e^{-t}C_4(e^{C_2 h} - e^{-h}).$$

Translating back to the language of semiflows, we see that $\varphi_h(z(t)) = z(t+h)$, and so

$$\sup_{0 \leq h \leq T} \|y(t+h) - \varphi_h(y(t))\| \leq e^{-t}C_4 \sup_{0 \leq h \leq T}(e^{C_2 h} - e^{-h}) = C_5 e^{-t}.$$

Therefore $y(t)$ is an asymptotic pseudotrajectory of the semiflow φ, and trajectories converge to the unique globally attracting fixed point (Proposition 21). Thus all trajectories of (5.3) converge to the point y^*, and the lemma is proved.

This lemma applies directly to the Q-learning ODE (5.2), and is useful for 2-player zero-sum games where there is a unique Nash distribution.

Lemma 49 *For 2-player zero-sum games, all trajectories of the ODE (5.2) will converge to the unique fixed point corresponding to a Nash distribution.*

PROOF Note that $(Q^1, Q^2) \in \mathbb{R}^{|A^1|+|A^2|}$. Since we consider 2-player zero-sum games, there is a matrix M such that $r^1(\cdot, \sigma^2) = M\sigma^2$ and $r^2(\cdot, \sigma^1) = -M^T\sigma^1$ (where M^T is the transpose of M). Define

$$S = \left\{ \begin{pmatrix} M & 0 \\ 0 & -M^T \end{pmatrix} x : x \in \mathbb{R}^{|A^1|+|A^2|} \right\}$$

$$D = \left\{ \begin{pmatrix} M & 0 \\ 0 & -M^T \end{pmatrix} \begin{pmatrix} \sigma^2 \\ \sigma^1 \end{pmatrix} : \sigma^i \in \Delta^i \right\}.$$

Since $\beta^i(Q^i) \in \text{int}(\Delta^i)$, it follows that $(r^1(\cdot, \beta^2(Q^2)), r^2(\cdot, \beta^2(Q^1))) \in \text{int}_S(D)$. Further, Q values in D are belief-based, so from Lemma 46 all trajectories with initial conditions in D converge to the Q values corresponding to the unique Nash distribution. Thus by Lemma 48 the result follows.

Proposition 50 *In 2-player zero-sum games, provided the Q values are bounded for all time, the strategies $\beta^i(Q^i_n)$ of individual Q-learners (5.1) will converge almost surely to the unique Nash distribution.*

PROOF Lemma 49 shows that all trajectories of the ODE (5.2) converge to the unique Nash distribution values. Standard stochastic approximation results (Section 1.3) show that the Q values will converge almost surely to the unique globally attracting fixed point of the ODE (5.2), corresponding to Nash distribution values. But the strategies of the players are simply the smooth best responses to these values, and so the strategies converge almost surely to the unique Nash distribution.

REMARK It is clear that a corresponding result can be obtained for 2-player partnership games. For Lemma 48 the presence of a unique globally attracting equilibrium in D is not particularly important: we have essentially shown that trajectories of (5.3) will converge to the chain-recurrent set of the semiflow restricted to S. A little extra work will show that this chain-recurrent set lies within D, and consists of Nash distributions (if these form an isolated set, i.e. if there are finitely or countably many of them).

REMARK A symmetric Q-learning equivalent of Model 1, Section 3.4, could be constructed, where now the population maintain Q values using (5.1), and the agent who plays the game selects an action according to a smooth best response to these values. This could be analysed in an identical manner to the N-player algorithm proposed here.

5.3 Multiple-timescales Q-learning

The individual Q-learning algorithm (5.1) can be studied using the smooth best response dynamics (1.8), and hence will not converge in the same situations as stochastic fictitious play or our simple actor–critic algorithm (3.5). In this section, we consider multiple-timescales Q-learning, in which each player learns at a different rate, in a similar fashion to the multiple-timescales actor–critic algorithm (4.6). The updates are exactly the same as for the individual Q-learning algorithm (5.1), except that the learning parameters are different for each player:

Multiple-timescales Q-learning algorithm

Each player i selects an action a_n^i using the strategy $\beta^i(Q_n^i)$, then updates Q_n^i according to

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \frac{\lambda_n^i \mathbb{I}_{\{a_n^i = a^i\}}}{\beta^i(Q_n^i)(a^i)} \left(R_n^i - Q_n^i(a^i)\right), \quad \text{for } a^i \in A^i, \tag{5.5}$$

where $\{\lambda_n^i\}_{n\geq 1}$ are deterministic sequences satisfying the standard conditions (1.15), and $\lambda_n^i/\lambda_n^j \to 0$ as $n \to \infty$ whenever $i < j$.

Following an identical argument as for individual Q-learning, it is clear that if $Q_n \to \tilde{Q}$ then $\beta^i(\tilde{Q}^i)$ gives Nash distribution strategies.

To analyse this algorithm in more detail, notice that for each i there is a separate timescale, with the associated ODE in each case being identical to that in the single-timescale case:

$$\dot{Q}^i(a^i) = r^i(a^i, \beta^{-i}(Q^{-i})) - Q^i(a^i).$$

For the multiple-timescales actor–critic algorithm (4.6) we had to assume that for fixed strategies $\pi^{\leq i}$ of the slow players, the fast players would converge to a unique joint smooth best response $B^{>i}(\pi^{\leq i})$. An exactly equivalent assumption needs to be made for the multiple-timescales Q-learning algorithm:

Assumption Q *For each $i = 2, \ldots, N-1$ there exists a Lipschitz function q^i such that $q^i(\pi^1, \ldots, \pi^{i-1})$ is the globally asymptotically stable equilibrium point of the ODE*

$$\dot{Q}^i = r^i\left(\cdot, [\pi^{<i}, B^{>i}[\pi^{<i}, \beta^i(Q^i)]]\right) - Q^i$$

where $B^{>i}$ is as defined in Section 4.2.

This assumption allows us to prove a result directly analogous to Theorem 42.

Theorem 51 *Under Assumption Q, and assuming the Q values remain bounded for all time, the values Q_n arising from the multiple-timescales Q-learning algo-*

rithm (5.5) satisfy

$$\left\|Q_n^i - r^i\left(\cdot, [\beta^1(Q_n^1), B^{>1}(\beta^1(Q_n^1))]\right)\right\| \to 0 \quad as \quad n \to \infty \quad a.s.$$

for $i = 2, \ldots, N$, and a suitable continuous time interpolation of the Q_n^1 is an asymptotic pseudotrajectory of the semiflow induced by the ODE

$$\dot{Q}^1 = r^1\left(\cdot, B^{>1}(\beta^1(Q^1))\right) - Q^1$$

PROOF Immediate from Theorem 40 and Assumption Q.

As noted for the multiple-timescales actor–critic algorithm, the convergence assumption (in this case Assumption Q) is vacuous in the case of 2-player games. From Theorem 51, we know that for the multiple-timescales Q-learning algorithm in a 2-player game,

$$\left\|Q^2 - r^2(\cdot, \beta^1(Q_n^1))\right\| \to 0 \quad \text{as} \quad n \to \infty,$$

and an interpolation of the Q_n^1 is an asymptotic pseudotrajectory of the semiflow induced by

$$\dot{Q}^1 = r^1(\cdot, \beta^2(\beta^1(Q^1))) - Q^1. \tag{5.6}$$

As in Section 5.2, this is most helpful in the case of zero-sum games, for which there is a unique fixed point.

Proposition 52 *In 2-player zero-sum normal form games, the strategies $\beta^i(Q_n^i)$ of multiple-timescales Q-learners (5.5) will converge almost surely to the unique Nash distribution, provided that the Q values remain bounded for all time.*

PROOF As for Lemma 49, it is a consequence of Lemma 48 and Proposition 43 that the dynamical system (5.6) has a globally attracting fixed point corresponding to the unique Nash distribution. Thus Theorem 51 and Proposition 21 show that the Q values converge to the Nash distribution values. Since the strategies are simply a continuous function of these Q values, the result follows.

As for the individual Q-learning algorithm (5.1), it is highly plausible that a similar result will hold for 2-player partnership games, but we do not show this here. Instead, consider a general game for which Assumption Q holds, but where Player 1 (the slowest) has only 2 actions.

Proposition 53 *Consider a game for which Assumption Q holds, where Player 1 has only 2 actions, and where there are finitely or countably many Nash distributions. If Player 1 uses Boltzmann action choice, and the Q values remain bounded for all time, then the strategies arising from the multiple-timescales Q-learning algorithm will converge to a Nash distribution.*

PROOF For this proof, to ease notation, write

$$\hat{r}(a) = r^1\left(a, B^{>1}(\beta^1(Q^1))\right), \quad a = 1, 2.$$

Since $\pi^1(2) = 1 - \pi^1(1)$, we see that $\hat{r}(a)$ depends on the scalar variable $\pi^1(1)$, which in turn depends on $Q^1(1)$ and $Q^1(2)$. Thus

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\hat{r}(a) &= \frac{\mathrm{d}\hat{r}(a)}{\mathrm{d}\pi^1(1)}\left(\frac{\partial\pi^1(1)}{\partial Q^1(1)}\dot{Q}^1(1) + \frac{\partial\pi^1(1)}{\partial Q^1(2)}\dot{Q}^1(2)\right) \\ &= \frac{\mathrm{d}\hat{r}(a)}{\mathrm{d}\pi^1(1)}\tau^{-1}\pi^1(1)\pi^1(2)\left(\dot{Q}^1(1) - \dot{Q}^1(2)\right), \end{aligned} \tag{5.7}$$

since

$$\pi^1(1) = \frac{1}{1 + \mathrm{e}^{\{Q^1(2) - Q^1(1)\}/\tau}} = 1 - \pi^1(2).$$

We wish to show that the quantity $U = Q^1(1) - Q^1(2)$ can be used to construct a Lyapunov function. This will be true if the time derivative $\dot{U} = \dot{Q}^1(1) - \dot{Q}^1(2)$ has fixed sign for all time, in which case either U or $-U$ is a Lyapunov function. But

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\dot{U} &= \tfrac{\mathrm{d}}{\mathrm{d}t}\left(\hat{r}(1) - Q^1(1) - \hat{r}(2) + Q^1(2)\right) \\ &= \tfrac{\mathrm{d}}{\mathrm{d}t}\left(\hat{r}(1) - \hat{r}(2)\right) - \dot{U} \\ &= \tau^{-1}\pi^1(1)\pi^1(2)\left(\dot{Q}^1(1) - \dot{Q}^1(2)\right)\left(\frac{\mathrm{d}\hat{r}(1)}{\mathrm{d}\pi^1(1)} - \frac{\mathrm{d}\hat{r}(2)}{\mathrm{d}\pi^1(1)}\right) - \dot{U} \\ &= \left\{\tau^{-1}\pi^1(1)\pi^1(2)\left(\frac{\mathrm{d}\hat{r}(1)}{\mathrm{d}\pi^1(1)} - \frac{\mathrm{d}\hat{r}(2)}{\mathrm{d}\pi^1(1)}\right) - 1\right\}\dot{U} \end{aligned}$$

and so $\dot{U}$ cannot change sign. Thus either U or $-U$ is a Lyapunov function, and by assumption there are finitely or countably many fixed points, so by Proposition 21 the result follows.

REMARK It is probable that this argument could be modified to show that for the same class of games the multiple-timescales actor–critic algorithm (4.6) will converge to Nash distribution, since the values $r^1(\cdot, \pi^{>1})$ under that model follow the same trajectories as the values Q^1 under multiple-timescales Q-learning.

REMARK Proposition 53 shows that multiple-timescales Q-learning will converge to the Nash distribution of our N-player matching pennies game (4.8).

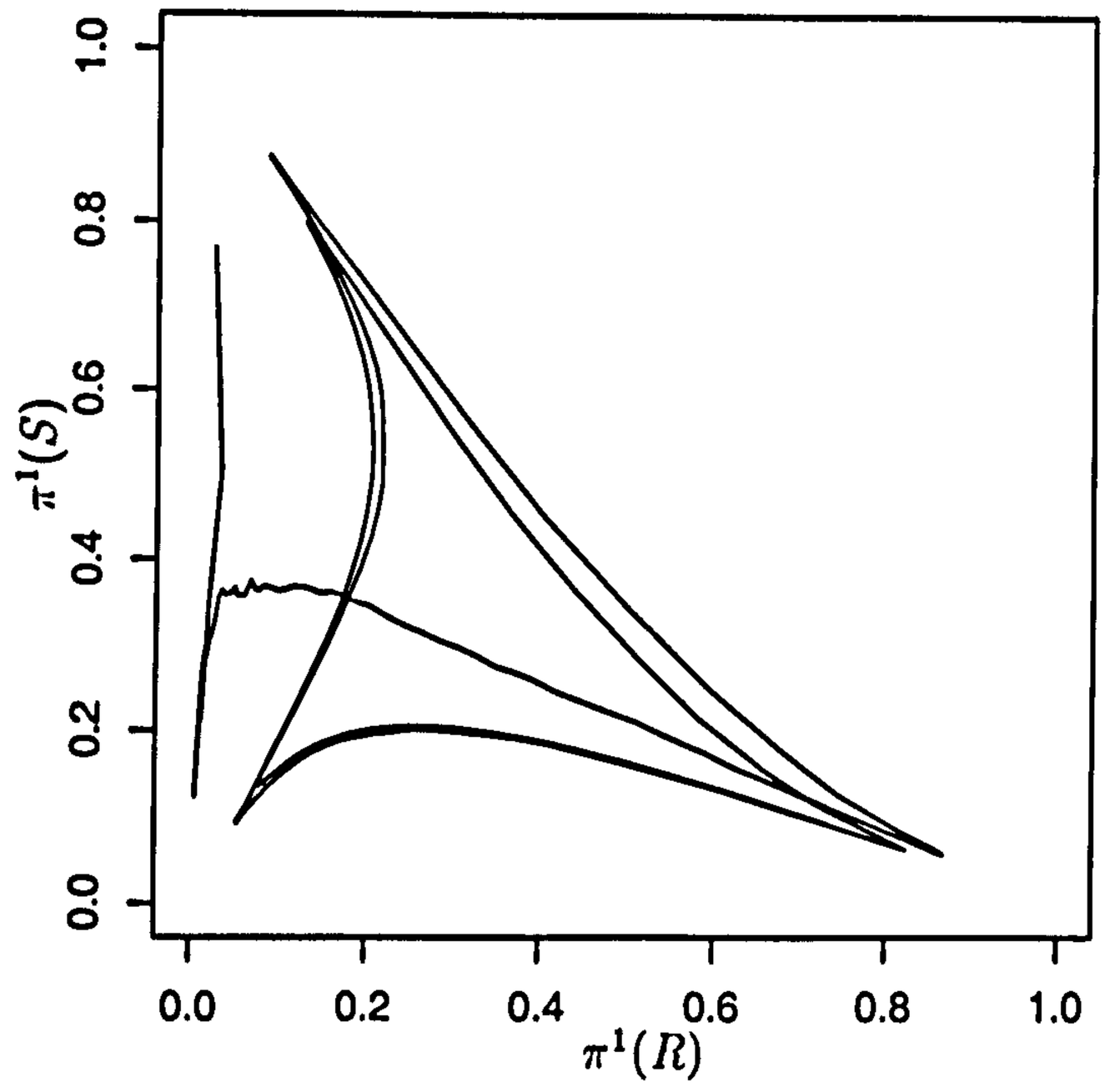
5.4 An example

In this section we consider Q-learning in Shapley's game (3.2). Note that for the smooth best response dynamics (1.8) there is an attracting limit cycle for the strategies, which implies that there is an attracting limit cycle for the Q-learning ODE (5.2). Therefore we would expect the Q values, and hence the strategies, arising from the individual Q-learning algorithm to cycle continuously.

On the other hand, there is a unique globally attracting fixed point of the perturbed smooth best response dynamics (4.7) for this game (Section 4.4.2). Therefore the analysis presented above for zero-sum games will carry through to show that the strategies arising from multiple-timescales Q-learning will converge to the unique Nash distribution in this case. See Fig. 5.1.

5.5 Conclusions

In this chapter we have dropped the explicit updating of a strategy, and instead consider an algorithm that is fully value-based (with the strategies played being simply a function of the values).



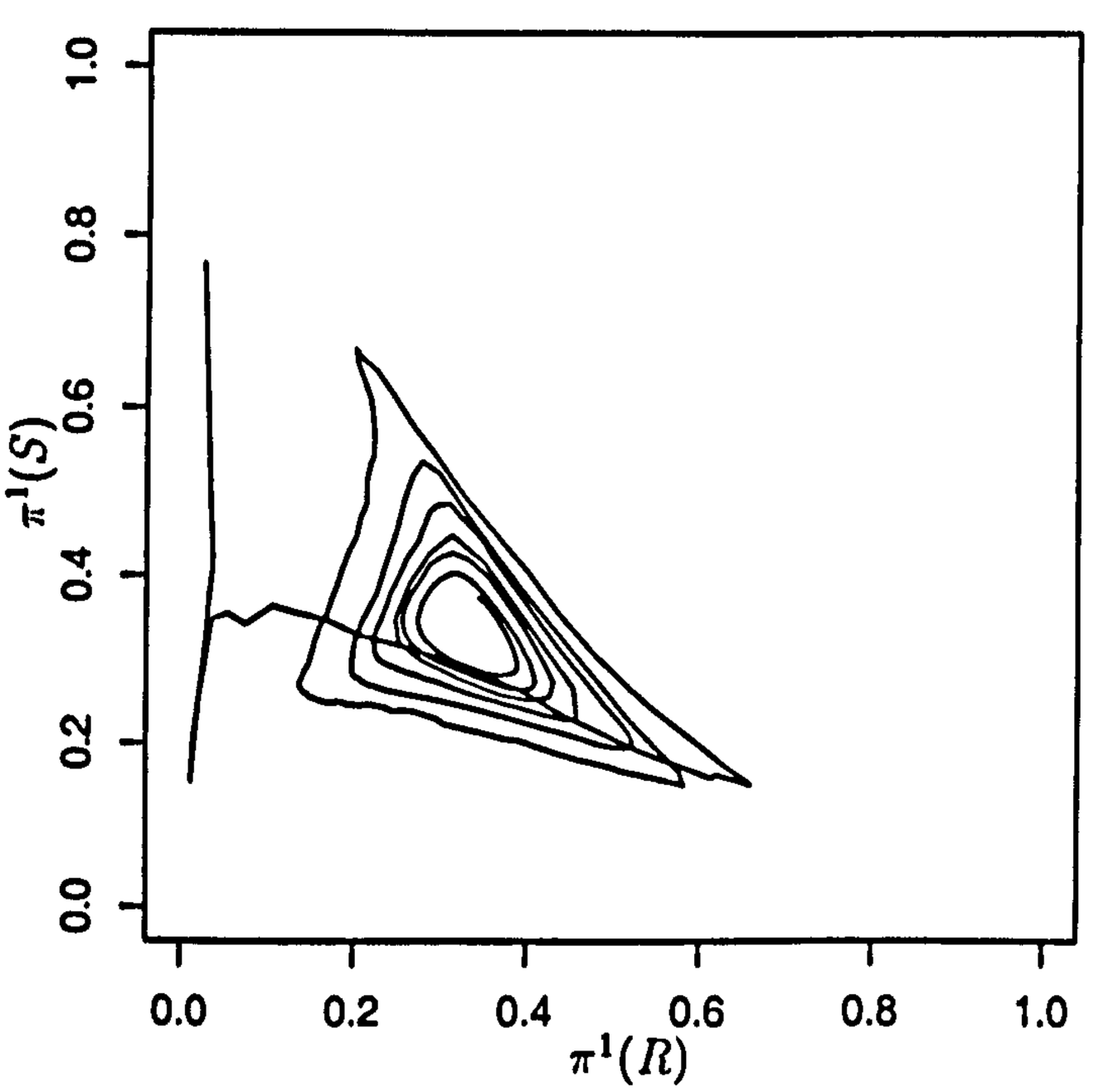


Figure 5.1: Strategies of Player 1 in Shapley's game (3.2) with Boltzmann action choices ($\tau = 0.1$) over 5×10^5 iterations of individual Q-learning with $\lambda_n = (n + 100)^{-0.9}$ (top) and of multiple-timescales Q-learning with $\lambda_n^1 = (n + 100)^{-0.9}$ and $\lambda_n^2 = (n + 100)^{-0.7}$ (bottom). For individual Q-learning the strategies follow a limit cycle, while for multiple-timescales Q-learning the strategies are converging towards the unique Nash distribution.

We have shown that under individual Q-learning the Q values are asymptotically belief-based, and therefore individual Q-learning behaves asymptotically in a similar manner to stochastic fictitious play and the actor–critic algorithm of Chapter 3, at least for 2-player games.

A multiple-timescales Q-learning algorithm has also been presented, which can be analysed in a similar manner to the multiple-timescales actor–critic algorithm. Again, a generic analysis of the behaviour of this multiple-timescales algorithm is not available, since the convergence assumption will not hold for all games.

The theoretical results have been verified by a numerical example, where we applied the algorithm in Shapley's game (3.2). As predicted, the strategies do not converge in the case of individual Q-learning, but do converge to the unique Nash distribution when multiple-timescales Q-learning is used.

Chapter 6

Weakened fictitious play and an actor–critic algorithm

So far we have considered players who choose smooth best responses, either to opponent play (in the case of stochastic fictitious play) or to estimates of action values (in the case of the reinforcement learning algorithms that have been considered). However, these algorithms result in agents that will play suboptimal actions with a non-vanishing probability even when an algorithm converges.

In this chapter we consider algorithms that are rational, in the sense of Bowling and Veloso (2002); if a rational algorithm converges to a fixed point then only actions with maximal expected reward will be played in the limit. Thus if these algorithms converge when applied in a game they must converge to the set of Nash equilibria (as opposed to the Nash distributions considered previously).

We study these algorithms by extending Hofbauer (1995) to show that the limit sets of fictitious play (Brown 1951), weakened fictitious play (Van der Genugten 2000), and a new actor–critic algorithm (to be introduced in Section 6.4) are each contained in the limit set of the best reply differential inclusion. This limit set is shown to coincide with the Nash equilibria of the game for 2-player zero-sum games, non-degenerate $2 \times m$ games and games solvable by strict dominance.

6.1 Discussion

Traditional analyses of discrete fictitious play, and its variants, have relied on *ad hoc* procedures to prove convergence (or otherwise) to the set of Nash equilibria, even though it has long been known that a fictitious play process is an Euler discretisation of the best response dynamics (Brown 1951). Robinson (1951) used "vector systems" to prove convergence in 2-player zero-sum games, an approach extended to allow consideration of a converging sequence of games (Vrieze and Tijs 1982) and "weakened fictitious play" (Van der Genugten 2000). On the other hand, Monderer and Shapley (1996) use a quasi-Lyapunov function approach in games with identical interests, and Milgrom and Roberts (1991) show directly that only serially undominated strategies will be played in the limit.

The methods of stochastic approximation used so far in this thesis will not apply in the case of fictitious play, since the best response dynamics (1.7) are not continuous. Some results (Delyon 1996; Tadić 1998) allow the stochastic approximation of discontinuous processes, but only where a Lyapunov function is present, which is not always the case for the situations we consider (though see Hofbauer and Sorin (2002) for indications of developments that may allow stochastic approximation of the best response differential inclusion)[1]. Further, the two-timescales approach used for our actor–critic algorithm is not easily studied using these methods.

Therefore we follow Hofbauer (1995) in using specific aspects of the proofs of stochastic approximation results; he shows that the limit set of a discrete fictitious play process is contained in the maximal invariant set of the best response dynamics, concentrating on symmetric 2-player games. This has been extended in several more recent works (Hofbauer and Sorin 2002; Berger 2003). Here we extend this work further, relating the limit sets of weakened fictitious play in N-player games to an invariant set of the best response differential inclusion.

[1] This since appeared as as Benaïm *et al.* (2003)

After discussing fictitious play models and the best response dynamics in Section 6.2, we prove in Section 6.3 that the limit set of a weakened fictitious play process is contained in an invariant set of the best response dynamics (it will be noted that a traditional fictitious play process, henceforth called a Brown–Robinson process, is a special case of a weakened fictitious play process). In Section 6.4 we introduce our actor–critic reinforcement learning algorithm, and show that it results in a weakened fictitious play process.

6.2 Fictitious play and the BR dynamics

It is clear that a Brown–Robinson process (1.6) is a decreasing step-size discretisation of the best response differential inclusion (1.7). It is intuitively obvious that, for sufficiently large n, a continuous-time interpolation of the Brown–Robinson process should track a trajectory of the best response dynamics (henceforth the BR dynamics). Indeed Hofbauer (1995) shows that any limit point of the Brown–Robinson process (1.6) will be contained in an invariant set of the BR dynamics (1.7).

However it has recently been observed (Van der Genugten 2000) that the speed of convergence of the Brown–Robinson process can be improved by considering weakened fictitious play, where players choose an ϵ_n-best response at stage n. Here player i's ϵ-best responses to opponent strategy π^{-i} are given by

$$\mathrm{BR}^i_\epsilon(\pi^{-i}) = \{b^i \in \Delta^i : r^i(b^i, \pi^{-i}) \geq \max_{\pi' \in \Delta^i} r^i(\pi^i, \pi^{-i}) - \epsilon\},$$

and we write

$$\mathrm{BR}_\epsilon(\pi) = \{(b^1, \ldots, b^N) \in \Delta^1 \times \cdots \times \Delta^N : b^i \in \mathrm{BR}^i_\epsilon(\pi^{-i}) \text{ for each } i\}. \qquad (6.1)$$

If $\epsilon_n \to 0$ as $n \to \infty$, so that players play asymptotically optimally, it is again intuitively obvious that the limiting behaviour of such processes is characterised by the limit behaviour of the BR dynamics (1.7), and we show that an identical

result can be proved for this process as is achieved by Hofbauer (1995) for the Brown-Robinson process.

Define a set A to be invariant for a differential inclusion if, for all $x_0 \in A$, there is a trajectory $x(t)$ that solves the differential inclusion, with $x(0) = x_0$ and $x(t) \in A$ for all t. We here state a theorem summarising previous results on the BR dynamics.

Theorem 54 *The maximal invariant set of the best response dynamics (1.7) is contained in the set of Nash equilibria for the following classes of games:*

1. *2-player zero-sum games,*

2. *non-degenerate 2-player $2 \times m$ games,*

3. *games solvable by iterated strict dominance.*

PROOF Class 1 is directly from Section 7 of Hofbauer (1995). Class 2 is the result of a recent paper (Berger 2003). Class 3 follows directly from the observation that the probability of playing a dominated strategy decreases exponentially for all time.

REMARK Following Hofbauer and Sigmund (1998) and Monderer and Shapley (1996) we note that the same results hold for games that are best response equivalent in mixed strategies to these games. In what follows we will not mention these rescalings, though clearly all our results will hold in this wider class.

REMARK Although it may seem natural that a similar result will hold for partnership games, this is not the case: despite the fact that all trajectories of the best response differential inclusion will converge to the set of Nash equilibria, an invariant set may contain other points.

6.3 Weakened fictitious play

In this section we extend Appendix B of Hofbauer (1995) to prove that limit points of a weakened fictitious play process will converge to an invariant set of the BR dynamics (1.7). This shows that a weakened fictitious play process will converge to the set of Nash equilibrium for the 4 classes of games listed in Theorem 54.

Consider a *continuous weakened fictitious play (CWFP) process* $f(t)$ with

$$\begin{aligned} f(t_n+\tau) &= \mathrm{e}^{-\tau} f(t_n) + (1-\mathrm{e}^{-\tau}) b(t_n) \quad \text{for} \quad 0 \le \tau \le \alpha_n = t_{n+1} - t_n \\ b(t_n) &\in \mathrm{BR}_{\epsilon_n}(f(t_n)) \\ t_n &\to \infty \end{aligned} \tag{6.2}$$

Any weakened fictitious play process (as described in Section 6.2), with belief vectors $\{\sigma_n\}_{n\ge 0}$, can be interpolated by a CWFP process with $t_n = \log(n)$ and $f(t_n) = \sigma_n$; however we consider a general CWFP process (6.2).

Definition 55 *A CWFP process (6.2) is called a* δ-path *if* $\alpha_n < \delta$ *and* $\epsilon_n < \delta$ *for all* n.

Thus for small δ the step sizes are small and the $b(t_n)$ are close to being best responses. Intuitively therefore, small δ should mean that the process (6.2) behaves similarly to the BR dynamics. This is what we will proceed to show.

Lemma 56 *Let* $\{\delta_i\}_{i\ge 1}$ *be a positive sequence of numbers such that* $\delta_i \to 0$ *as* $i \to \infty$, *and let* f_i *be a* δ_i*-path for each* i. *Then for given* T *there is a subsequence* δ_{i_k} *such that the functions* $f_{i_k}(t)$ *converge uniformly on the interval* $[0,T]$ *as* $k \to \infty$. *Any such limit* $\pi(t)$ *is a solution of the BR dynamics (1.7) on* $[0,T]$.

Proof We closely follow the proof of Hofbauer (1995). Let $t_n^{(i)}$ denote the 'interpolation points' of f_i. Note that (6.2) is equivalent to

$$\tfrac{\mathrm{d}}{\mathrm{d}\tau} f(t_n^{(i)}+\tau) = b(t_n^{(i)}) - f(t_n^{(i)}+\tau) \quad \text{for} \quad 0 \le \tau \le \alpha_n^{(i)} = t_{n+1}^{(i)} - t_n^{(i)} \tag{6.3}$$

and an integrating factor shows that f_i satisfies

$$f_i(t) = e^{-t} f_i(0) + \int_0^t e^{s-t} b_i(s)\, ds,$$

where b_i is a *cad lag* step function with values $b_i(t_n^{(i)}) \in BR(f_i(t_n^{(i)}))$.

From the differential equation representation (6.3) it follows that the f_i are uniformly Lipschitz continuous, and thus by the Arzela-Ascoli theorem there is a subsequence δ_{i_k} such that the functions f_{i_k} are uniformly convergent on $[0, T]$; suppose that $f_{i_k} \to \pi$.

Now consider the functions b_{i_k}. Since we are working in compact spaces, there exists a weak accumulation point. Take any such accumulation point b, and consider the value of $b(t)$ for arbitrary $t \in [0, T]$. If we can show that $b(t) \in BR(\pi(t))$ we are done, since then π satisfies

$$\pi(t) = e^{-t}\pi(0) + \int_0^t e^{s-t} b(s)\, ds, \quad b(s) \in BR(\pi(s)),$$

which is the integral representation of the BR dynamics (1.7).

Fix t and choose $\eta > 0$. In what follows let B be the open unit ball in $\mathbb{R}^{|A^1|+\cdots+|A^N|}$, and for $x \in \mathbb{R}^{|A^1|+\cdots+|A^N|}$ let $x + \gamma B = \{x + \gamma z : z \in B\}$. Similarly, for sets $X, Y \subset \mathbb{R}^{|A^1|+\cdots+|A^N|}$ let $X + Y = \{x + y : x \in X, y \in Y\}$. By the definition (6.1) of ϵ-best responses, and the continuity of BR_ϵ at $\epsilon = 0$, we can choose $\eta_1, \eta_2 > 0$ such that

$$\begin{aligned} \delta < \eta_1 \quad &\Rightarrow \quad BR_\delta(\pi(t)) \subset BR(\pi(t)) + \tfrac{\eta}{2} B, \quad \text{and} \\ z \in \pi(t) + \eta_2 B \quad &\Rightarrow \quad BR_\delta(z) \subset BR_\delta(\pi(t)) + \tfrac{\eta}{2} B. \end{aligned}$$

Since $f_{i_k} \to \pi$, and the f_{i_k} are uniformly Lipschitz, we can choose $K > 0$ such that, for all $k \geq K$,

$$\begin{aligned} &\delta_{i_k} < \eta_1, \quad \text{and} \\ &\|f_{i_k}(t - t') - \pi(t))\| < \eta_2 \quad \text{for all } t' < \delta_{i_k}. \end{aligned}$$

Then, since $b_{i_k}(t) \in BR(f_{i_k}(t - t'))$ for some $t' < \delta_{i_k}$, for $k \geq K$,

$$b_{i_k}(t) \in BR_{\delta_{i_k}}(f_{i_k}(t - t')) \subset BR_{\delta_{i_k}}(\pi(t)) + \tfrac{\eta}{2} B \subset BR(\pi(t)) + \eta B.$$

Thus $b(t) \in \mathrm{BR}(x(t))$, and the proof is complete.

We use this lemma to consider the limit sets of CWFP process (6.2) for which $\alpha_n \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$. As already noted, this case includes the interpolation of a discrete weakened fictitious play process.

Lemma 57 *Consider a CWFP process f with $\alpha_n \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$. The set of limit points of this process is an invariant set for the BR dynamics (1.7).*

PROOF This proof is in Appendix B of Hofbauer (1995). We include it here for completeness.

Let L be the set of limit points of f (which exists because $f(t)$ is in a compact space). L is a limit set, and so is compact; if for arbitrary $p \in L$ and $T > 0$ we can construct a solution $x(t) \in L$, $-T \leq t \leq T$ with $x(0) = p$ then the result follows.

For any $p \in L$ there exists a sequence $\{s_i\}_{i \geq 0}$ with $s_i \to \infty$ as $i \to \infty$ such that $f(s_i) \to p$. Define functions $x_i(t) = f(s_i + t)$ for $-T \leq t \leq T$ (since $s_i \to \infty$ we can assume w.l.o.g. that $s_i > T$), and define $f_i(t) = x_i(t - T)$ for $0 \leq t \leq 2T$. Then the $f_i(t)$ satisfy the conditions of Lemma 56, and there exists an accumulation point $g(t)$, $0 \leq t \leq 2T$, which is a solution of the BR dynamics (1.7). Defining $x(t) = g(t + T)$, $-T \leq t \leq T$, it is obvious that x is an accumulation point of the x_i and satisfies the BR dynamics. From the definition of the x_i it follows that $x(0) = p$, and since x is an accumulation point of the x_i it follows that $x(t) \in L$ for $-T \leq t \leq T$.

Theorem 58 *The limit set of any weakened fictitious play process is contained in the maximal invariant set of the BR dynamics.*

PROOF As noted at the start of this section, any weakened fictitious process has an interpolation (6.2) with $t_n = \log(n)$. Thus $\alpha_n = \log(n+1) - \log(n) \approx 1/n$. Since the limit set of the discrete process is clearly contained in the limit set of the interpolation, the result follows immediately from Lemma 57.

Corollary 59 *The limit set of the Brown-Robinson process is contained in the maximal invariant set of the BR dynamics.*

PROOF The Brown-Robinson process is a special case of weakened fictitious play.

Corollary 60 *Any weakened fictitious play process will converge to the set of Nash equilibria in 2-player zero-sum games, non-degenerate $2 \times m$ games, and games solvable by iterated strict dominance.*

PROOF This is immediate from Theorems 54 and 58.

REMARK It is clear that the result of a fictitious play process where the estimates of action values $r^i(a^i, \pi_n^{-i})$ are made using a converging sequence of reward functions r_n^i is a weakened fictitious play process, since for any $\epsilon > 0$, we can choose $\delta > 0$ such that

$$\|r_n^i - r^i\|_\infty < \delta \Rightarrow \operatorname*{argmax}_{\pi^i} r_n^i(\pi^i, \pi^{-i}) \subset \mathrm{BR}_\epsilon^i(\pi^{-i}).$$

REMARK We show that the type of fictitious play in converging 2-player zero-sum games studied by Vrieze and Tijs (1982) can easily be incorporated in the framework of weakened fictitious play. Vrieze and Tijs (1982) consider estimates $U_n^i(a^i)$ that are updated according to

$$U_{n+1}^i(a^i) = U_n^i(a^i) + r_n^i(a^i, a_n^{-i}).$$

If $r_n^i = r^i$ for all n then this is exactly equivalent to the fictitious play process described above, with $r^i(a^i, \sigma_n) = U_n^i(a^i)/n$, but when we take $r_n^i \to r^i$ as $n \to \infty$ the equivalence is not immediate. However, fix $\epsilon > 0$ and let $n(\epsilon)$ be such that $\|r_n^i - r^i\|_\infty < \epsilon$ for all $n \geq n(\epsilon)$. For $n \geq n(\epsilon)$ we see that

$$U_n^i(a^i) = U_{n(\epsilon)}^i(a^i) + \sum_{k=n(\epsilon)+1}^{n} r_n^i(a^i, a_n^{-i})$$

and so

$$\left| U_n^i(a^i) - \{U_{n(\epsilon)}^i(a^i) + \sum_{k=n(\epsilon)+1}^{n} r^i(a^i, a_n^{-i})\} \right| < (n - n(\epsilon))\epsilon < n\epsilon.$$

But $\sum_{k=n(\epsilon)+1}^{n} r^i(a^i, a_n^{-i}) = nr^i(a^i, \sigma_n^{-i}) - n(\epsilon)r^i(a^i, \sigma_{n(\epsilon)}^{-i})$ and so

$$|U_n^i(a^i)/n - r^i(a^i, \sigma_n^{-i})| < \left| \frac{U_{n(\epsilon)}^i(a^i)}{n} - \frac{n(\epsilon)}{n} r^i(a^i, \sigma_{n(\epsilon)}^{-i}) \right| + \epsilon$$

$$< 2\epsilon \quad \text{for sufficiently large } n.$$

Thus optimal actions under U_n^i are ϵ_n-optimal actions under $r^i(a^i, \sigma_n^{-i})$, with $\epsilon_n \to 0$ as $n \to \infty$; Vriezc and Tijs's fictitious play for converging games is a weakened fictitious play of thc limit game. Note that this also shows that the weakened fictitious play of Van der Genugten (2000), which is based on Vrieze and Tijs's fictitious play in converging games, is also a weakened fictitious play of the limit gamc, and so the tight conditions placed on ϵ_n by Van der Genugten (2000) can bc rclaxcd to general $\epsilon_n \to 0$.

REMARK Since thc only conditions we place on the α_n are that $\alpha_n \to 0$ and $\sum_{n\geq 1} \alpha_n = \infty$ this allows the consideration of fictitious play processes where beliefs about opponent strategics places more weight on recent observations. Therefore define *generalised weakened fictitious play* to be any process

$$\sigma_{n+1} = (1 - \lambda_n)\sigma_n + \lambda_n \mathrm{BR}_{\epsilon_n}(\sigma_n),$$

where $\lambda_n \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$, and $\sum_{n\geq 1} \lambda_n = \infty$. The interpolation (6.2) requires that $\alpha_n = -\log(1 - \lambda_n)$, and the condition $\alpha_n \to 0$ corresponds to thc condition $\lambda_n \to 0$. Thus a generalisecd weakened fictitious play process will converge to thc maximal invariant set of the BR dynamics. This may seem counter-intuitive, since we have relied on the condition $\sum_{n\geq 1} {\lambda_n}^2 < \infty$ throughout this thesis. However, recall that this is essentially a condition to bound the variance, and since there is no (explicit) stochasticity involved in weakened fictitious play the condition is no longer nccessary.

6.4 An actor–critic learning algorithm

We now introduce an actor–critic reinforcement learning algorithm, which we demonstrate will result in a weakened fictitious play process. In common with the other algorithms introduced in this thesis, this actor–critic process does not require players to observe the actions or rewards of the other players, and is just as applicable in a single-agent learning task. Indeed, this algorithm is also rational, in the sense of Bowling and Veloso (2002), and will converge to the set of optimal actions in a single-agent problem.

The algorithm of this chapter is a modification of that studied in Chapter 3, in which players adjust their strategies towards best responses (instead of smooth best responses). If, as in that chapter, the estimates Q_n satisfy

$$\|Q_n - \mathfrak{L}(\pi_n)\| \to 0 \quad \text{as} \quad n \to \infty$$

then the π_n will be a (generalised) weakened fictitious play process, and the results of the previous section will hold.

However the best responses remove the opportunity to use the full strength of Borkar's results (Theorem 23), which require continuity. Also, since strategies no longer remain completely mixed for all time, there are added complications in checking that all actions are updated infinitely often so that the Q values are asymptotically accurate. We start by giving conditions under which all actions will be played infinitely often:

Lemma 61 *Suppose that at time n, the probability of playing each action is bounded below by $\zeta_n = (C+n)^{-\rho}$ for some $C > 0$ and $\rho \in (0,1)$. Then with probability 1 all actions will be played infinitely often.*

PROOF This is a simple consequence of the Borel-Cantelli lemma.

Although this condition on the strategies π_n will not arise naturally, it can be enforced by having player i select actions according to a projection $\psi_{\zeta_n}(\pi_n^i)$ of the

current strategy π_n^i, where

$$\psi_\zeta(\pi) = (1-s)\pi^i + s\underline{1}, \quad s = \max\left\{0, \max_{a\in A} \frac{\zeta - \pi(a)}{1-\pi(a)}\right\}.$$

This is a projection into the space where all strategies are played with sufficiently high probability. An alternative would be to simply choose a random action with probability ζ_n and otherwise use π_n^i, but this disturbs the strategies unnecessarily if they are all played with sufficiently high probability anyway.

We can now specify our algorithm:

Discontinuous actor–critic algorithm

Each player i selects an action a_n^i using the strategy $\psi_{\zeta_n}(\pi_n^i)$, then updates π_n^i and Q_n^i according to

$$\begin{aligned} \pi_{n+1}^i &= (1-\mu_{n+1})\pi_n^i + \mu_{n+1} b^i(Q_n^i) \\ Q_{n+1}^i(a^i) &= Q_n^i(a^i) + \lambda_{n+1}^i \mathbb{I}_{\{a_n^i = a^i\}}(R_n^i - Q_n^i(a^i)), \quad \text{for } a^i \in A^i, \end{aligned} \tag{6.4}$$

where:

- $b^i(Q_n^i) \in \operatorname{argmax}_{\pi\in\Delta^i} \pi \cdot Q_n^i$,
- $\mu_{n+1} = (C_\mu + n)^{-\rho_\mu}$ for some $C_\mu > 0$ and $\rho_\mu \in (0.5, 1]$,
- $\lambda_{n+1}^i = (C_\lambda + c_n^i(a_n^i))^{-\rho_\lambda}$ for some $C_\lambda > 0$ and $\rho_\lambda \in (0.5, \rho_\mu)$, where $c_n^i(a^i) = \sum_{k=0}^n \mathbb{I}_{\{a_n^i = a^i\}}$.
- $\zeta_n = (C_\zeta + n)^{-\rho_\zeta}$ for some $C_\zeta > 0$ and $\rho_\zeta \in (0, \rho_\mu - \rho_\lambda)$.

We start by showing that the Q values are asymptotically accurate estimates of the action values.

Lemma 62 *Fix i, a^i, and let $\{\nu_k\}_{k\geq 1}$ be the sequence of times when action a^i is played by player i. Define the differences*

$$D_k = Q_{\nu_k}^i(a^i) - r^i(a^i, \pi_{\nu_k}^{-i}).$$

Then $D_k \to 0$ *almost surely.*

PROOF First notice that ν_k is well-defined for all k, by Lemma 61. Taking expectations with respect to the history up to ν_k,

$$\begin{aligned}\mathbb{E}[D_{k+1} - D_k] =& \mathbb{E}[Q^i_{\nu_{(k+1)}}(a^i) - r^i(a^i, \pi^{-i}_{\nu_{(k+1)}}) - Q^i_{\nu_k}(a^i) + r^i(a^i, \pi^{-i}_{\nu_k})] \\ =& \mathbb{E}[Q^i_{(\nu_k)+1}(a^i) - Q^i_{\nu_k}(a^i)] - \mathbb{E}[r^i(a^i, \pi^{-i}_{\nu_{(k+1)}}) - r^i(a^i, \pi^{-i}_{\nu_k})] \\ =& \lambda^i_{(\nu_k)+1}\left\{r^i(a^i, \pi^{-i}_{\nu_k}) - Q^i_{\nu_k}(a^i) + [r^i(a^i, \psi_{\zeta_{\nu_k}}(\pi_{\nu_k})^{-i}) - r^i(a^i, \pi^{-i}_{\nu_k})]\right. \\ & \left. - \frac{\mathbb{E}[r^i(a^i, \pi^{-i}_{\nu_{(k+1)}}) - r^i(a^i, \pi^{-i}_{\nu_k})]}{\lambda^i_{(\nu_k)+1}}\right\}\end{aligned}$$

and so

$$D_{k+1} = (1 - \lambda^i_{(\nu_k)+1})D_k + \lambda^i_{(\nu_k)+1}(M_k + F_k - E_k),$$

where M_k is a bounded martingale difference, and

$$F_k = r^i(a^i, \psi_{\zeta_{\nu_k}}(\pi_{\nu_k})^{-i}) - r^i(a^i, \pi^{-i}_{\nu_k})$$

$$E_k = \{r^i(a^i, \pi^{-i}_{\nu_{k+1}}) - r^i(a^i, \pi^{-i}_{\nu_k})\}/\lambda^i_{(\nu_k)+1}.$$

If we can show that $\|F_k\| \to 0$ and $\|E_k\| \to 0$ almost surely, then Lemma 1 of Singh *et al.* (2000) shows that $D_k \to 0$ almost surely. This is trivial for F_k, since $\zeta_n \to 0$ and r^i is continuous. For E_k, notice that

$$\|r^i(a^i, \pi^{-i}_{n+1}) - r^i(a^i, \pi^{-i}_n)\|_\infty \leq C\mu_{n+1}$$

for some C (depending only on the reward function r^i), and so

$$C^{-1}\|E_k\| \leq \frac{\sum_{j=\nu_k}^{\nu_{k+1}-1} \mu_{j+1}}{\lambda^i_{(\nu_k)+1}} \leq \frac{(\nu_{k+1} - \nu_k)\mu_{(\nu_k)+1}}{\lambda^i_{(\nu_k)+1}} \leq C' \frac{\nu_{k+1} - \nu_k}{\nu_k{}^{\rho_\mu - \rho_\lambda}},$$

where C' depends on the constants C_μ and C_λ. Thus, by the Borel-Cantelli lemma, we see that $\|E_k\| \to 0$ almost surely if, for arbitrary $\delta > 0$,

$$\sum_{k \geq 1} \mathbb{P}\left(\frac{\nu_{k+1} - \nu_k}{\nu_k{}^{\rho_\mu - \rho_\lambda}} > \delta\right) < \infty. \tag{6.5}$$

Let j be the greatest integer less than or equal to $\delta(\nu_k)^{\rho_\mu - \rho_\lambda}$. Then

$$\begin{aligned}\mathbb{P}\left(\frac{\nu_{k+1}-\nu_k}{{\nu_k}^{\rho_\mu-\rho_\lambda}} > \delta\right) &= \mathbb{P}\left(\nu_{k+1} > \nu_k + j\right) \\ &\leq (1-\zeta_{\nu_k+1})(1-\zeta_{\nu_k+2})\cdots(1-\zeta_{\nu_k+j}) \\ &< (1-\zeta_{\nu_k+j})^j \\ &< \exp\left\{-j\zeta_{\nu_k+j}\right\}.\end{aligned}$$

It follows from the definition of j that

$$\begin{aligned}\mathbb{P}\left(\frac{\nu_{k+1}-\nu_k}{{\nu_k}^{\rho_\mu-\rho_\lambda}} > \delta\right) &< \exp\left\{-\left(\delta(\nu_k)^{\rho_\mu-\rho_\lambda}-1\right)\zeta_{\nu_k+j}\right\} \\ &= \exp\zeta_{\nu_k+j} \times \exp\left\{-\delta(\nu_k)^{\rho_\mu-\rho_\lambda}\zeta_{\nu_k+j}\right\}.\end{aligned}$$

Since ζ_n is decreasing with n, it follows that $\exp\zeta_{\nu_k+j} < C_1$ for some constant $C_1 > 0$, and additionally from the definition of j and ζ_{ν_k+j} we see that

$$\begin{aligned}\zeta_{\nu_k+j} &= (C_\zeta + \nu_k + j)^{-\rho_\zeta} \\ &> (C_\zeta + \nu_k + \delta(\nu_k)^{\rho_\mu-\rho_\lambda})^{-\rho_\zeta} \\ &= (\nu_k)^{-\rho_\zeta}\left(C_\zeta(\nu_k)^{-1} + 1 + \delta(\nu_k)^{\rho_\mu-\rho_\lambda-1}\right)^{-\rho_\zeta} \\ &> C_2(\nu_k)^{-\rho_\zeta} \quad \text{for some } C_2 > 0.\end{aligned}$$

Therefore, it follows that

$$\mathbb{P}\left(\frac{\nu_{k+1}-\nu_k}{{\nu_k}^{\rho_\mu-\rho_\lambda}} > \delta\right) < C_1\exp\left\{-\delta C_2(\nu_k)^{\rho_\mu-\rho_\lambda-\rho_\zeta}\right\}$$

Since $\nu_k \geq k$, this is bounded above by $C_3\exp\{-C_4k^{\rho_\mu-\rho_\lambda-\rho_\zeta}\}$ for some $C_3, C_4 > 0$.

Now, for $\eta > 0$,

$$\int_0^\infty \mathrm{e}^{-C_4x^{-\eta}}\mathrm{d}x \overset{y:=x^{-\eta}}{=} \eta^{-1}\int_0^\infty y^{\eta^{-1}-1}e^{-C_4y}\mathrm{d}y = \eta^{-1}\Gamma(\eta^{-1}){C_4}^{-\eta^{-1}}$$

where Γ is the Gamma function. Therefore $\sum_{k\geq 0} C_3\exp\{-C_4k^{\rho_\mu-\rho_\lambda-\rho_\zeta}\} < \infty$, since $\rho_\zeta < \rho_\mu - \rho_\lambda$, and we see that (6.5) holds.

This shows that at the times that an action is played, its Q value is close to the action's current reward. However, since all actions are played with strictly

positive probability at all times this shows that the Q values are asymptotically accurate, as required, and so the two-timescales nature of the continuous actor-critic algorithm (3.5) passes through to this discontinuous version. Note that the careful choice of ζ_n is vital to ensure that actions are played sufficiently frequently to gain this asymptotic accuracy.

Theorem 63 *The π_n resulting from the discontinuous actor-critic algorithm (6.4) are a generalised weakened fictitious play process.*

PROOF By Lemma 62, and the fact that any action is played with strictly positive probability at any time, $\|Q_n^i(a^i) - r^i(a^i, \pi_n^{-i})\|_\infty \to 0$ as $n \to \infty$. Therefore $b^i(Q_n^i) \in \mathrm{BR}_{\epsilon_n}^i(\pi_n^{-i})$ for some $\epsilon_n \to 0$. Thus the π_n are a generalised weakened fictitious play process.

Corollary 64 *The strategies arising from the discontinuous actor-critic algorithm (6.4) will converge to the set of Nash equilibria in 2-player zero-sum games, non-degenerate $2 \times m$ games, and games solvable by iterated strict dominance.*

PROOF This is immediate from Theorem 63 and Corollary 60.

We conclude our analysis of the discontinuous actor-critic algorithm by presenting the results of an experiment using the 2-player matching pennies game, with reward matrix

$$\begin{pmatrix} (1,0) & (0,1) \\ (0,1) & (1,0) \end{pmatrix}$$

This is a constant-sum game, so strategies evolve exactly as if it is zero-sum (Hofbauer and Sigmund 1998), and should therefore converge to the unique Nash equilibrium where each player plays each action with probability 1/2. In Fig. 6.1 we see that even after 5×10^6 iterations the strategies are still cycling, but are spiralling towards the Nash equilibrium.

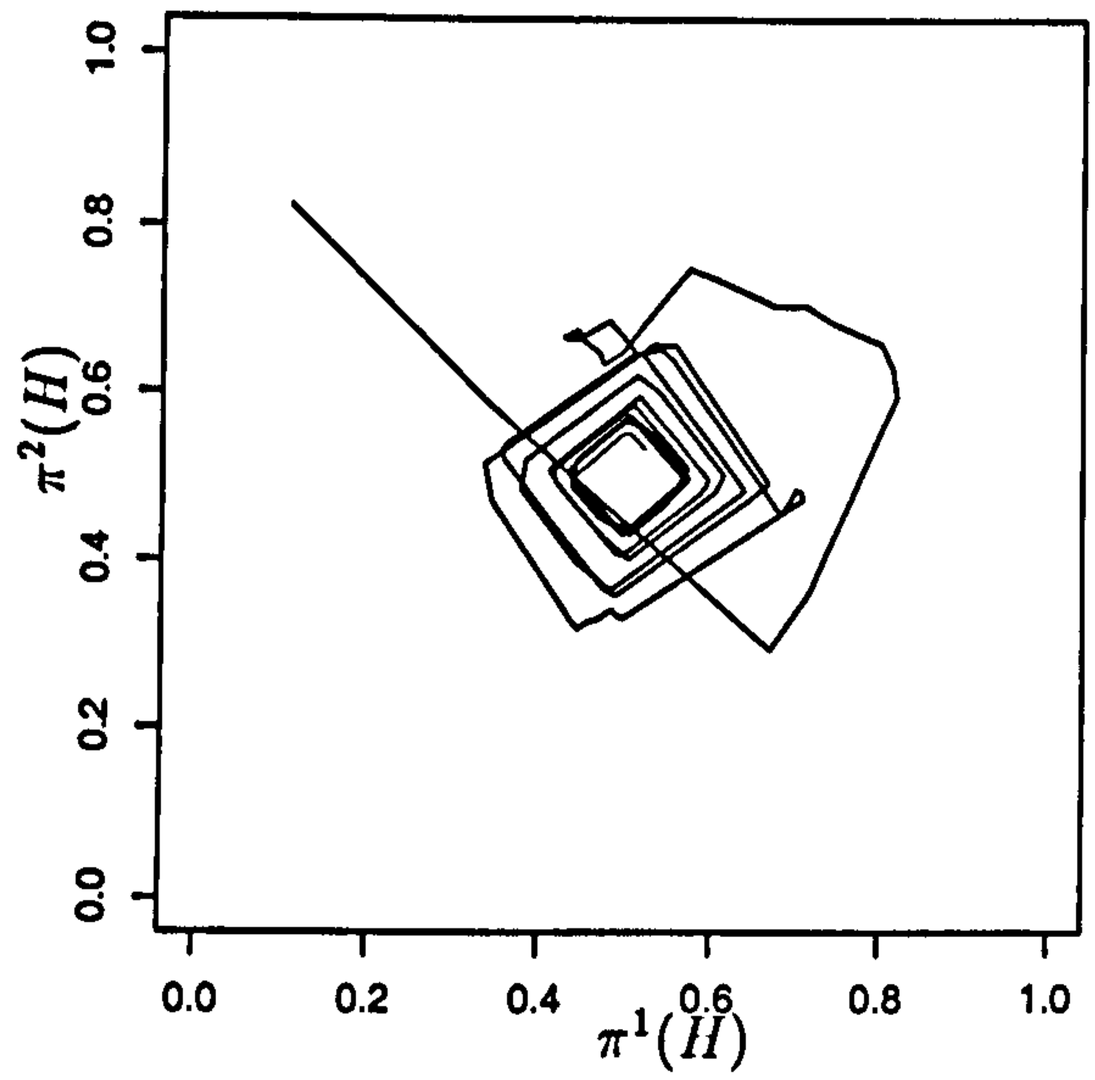


Figure 6.1: Strategies over 5×10^6 iterations of the discontinuous actor–critic algorithm in the 2-player matching pennies game. The parameters are $\rho_\mu = 1.0$, $\rho_\lambda = 0.6$, and $\rho_\zeta = 0.1$. After an initial period the strategies spiral clockwise towards the equilibrium point.

6.5 Conclusion

In this chapter we returned to actor–critic algorithms, and investigated what happens when best responses are used instead of smooth best responses. We studied this algorithm by generalising results on fictitious play, relating weakened fictitious play processes to the best response dynamics (1.7).

In particular, we have shown that the limit set of any generalised weakened fictitious play process is contained in the maximal invariant set of the BR dynamics (1.7). Using previous results about these dynamics, this shows that a generalised weakened fictitious play process will converge to the set of Nash distributions for 2-player zero-sum games, non-degenerate $2 \times m$ games, and games solvable by iterated strict dominance. This generalises a result of Van der Genugten (2000), who placed tight conditions on the sequence $\{\epsilon_n\}$ to prove convergence of values in 2-player zero-sum games.

We then showed that an actor–critic algorithm, similar to that of Chapter 3, results in a generalised weakened fictitious play process, and therefore converges to equilibrium in the same class of games. However, the fact that play can converge to strategies where not all actions are played with positive probability means that similar problems might be faced as with the simple learning model of Chapter 2 (i.e. convergence to non-Nash pure strategy combinations). However, we showed that the Q values are asymptotically accurate if the strategies played are carefully controlled, and this means that convergence to a non-Nash point is not possible.

Chapter 7

Smooth best responses in stochastic games

The reinforcement learning algorithms studied in this thesis were inspired by similar algorithms applied in Markov decision processes. It therefore seems reasonable to expect that learning to play stochastic games is within the reach of extensions of our algorithms, since stochastic games are simply a combination of Markov decision processes and normal form games.

While convergence results have so far proved elusive in this area, progress has been made in the study of smooth best responses. This is related to an observation of John (1994) that Q-learning under persistent exploration does not take into account the fact that suboptimal actions, which are potentially very costly, will always be played (since for Bellman's equations $V(x) = \max_{a \in A(x)} Q(x,a)$). John's suggestion is to take into account the fact that the policy $\pi(x)$ played at state x is a function of the Q values at state x, so that writing $\pi(x,a)$ for the probability of playing action a at state x, the Q value formulation of Bellman's equations becomes

$$\begin{aligned} Q(x,a) &= r(x,a) + \delta \sum_{y \in X} P_{xy}(a) V(y) \\ V(x) &= \sum_{a \in A(X)} \pi(x,a) Q(x,a), \qquad \text{for all } x \text{ and } a. \end{aligned} \tag{7.1}$$

Now $V(x)$ represents the 'expected Q value' at state x, and $Q(x,a)$ gives the actual expected future discounted reward if action a is played and then policies $\pi(x)$ are followed (as opposed to the maximal expected future discounted reward, which is obtained when $V(x) = \max_{a \in A(x)} Q(x,a)$). Under the kind of learning algorithm we have been studying (other than in Chapter 6), smooth best responses will be played for all time, and so again if the Q values (and V values) are to represent the expected future reward to be received we must use $V(x) = \sum_{a \in A(x)} \pi(x,a) Q(x,a)$.

However Littman (1996) observed that if Boltzmann smooth best responses are played there may not even be a unique solution to the modified Bellman equations (7.1). This is because the map $Q \mapsto V$ with Boltzmann smooth best responses is not contracting (with respect to the L_∞ norm). We are therefore interested in smooth best responses π such that this map is non-expansive.

Two methods of choosing smooth best responses have been proposed: McNamara *et al.* (1997) use a model where players make errors, and the probability of playing a suboptimal action is a function of the loss induced by doing so, whereas Fudenberg and Kreps (1993), based on Harsanyi (1973), suggest that players add a random perturbation to their action values before selecting an action to maximise the perturbed values. We show that non-expansive best responses are impossible with the former method, but provide sufficient conditions and a non-expansive smooth best response function based on Fudenberg and Kreps' model.

7.1 Error-based methods are expansive

Under the intuitive scheme of McNamara *et al.* (1997), the representation of $V(x)$ in (7.1) becomes

$$V(x) = \frac{\sum_{a \in A} h(c(x,a)) Q(x,a)}{\sum_{a \in A} h(c(x,a))}$$

where $c(x,a) = \max_{b \in A} Q(x,b) - Q(x,a)$ and h is a function satisfying the following properties:

- $h(c) > 0$ for all $c \geq 0$,
- $h'(c) \leq 0$ for all $c \geq 0$,
- $h(c) \to 0$ as $c \to \infty$.

These conditions say that the probability of playing any action is positive ($h(c) > 0$), but that the probability of playing an action decreases towards 0 as the cost $c(x, a)$ of playing that action gets large ($h'(c) \leq 0$ and $h(c) \to 0$ as $c \to \infty$).

Since it is required that the smooth best response be non-expansive at all states x, in the sequel we will fix x and write $q_a = Q(x, a)$, $c_a = c(x, a)$. We will derive additional constraints on h that ensure the map $q \to V(q)$ is a non-expansion, where $q \in \mathbb{R}^m$, $V(q) = \frac{\sum_a q_a h(c_a)}{\sum_a h(c_a)}$, and m, the number of actions, is arbitrary. The set of constraints on h will then prove to be inconsistent, showing that no suitable h exists.

Suppose one of the partial derivatives $\frac{\partial V(q)}{\partial q_a}$ is negative. We see that for some small $\delta \in \mathbb{R}^+$

$$\frac{\partial V(q)}{\partial q_a} < 0 \quad \Rightarrow \quad V(q + \delta_a) < V(q)$$

where δ_a is a vector with 0 in every entry except for a δ in position a. Then,

$$V(q + 2\underline{\delta}) = V(q) + 2\delta > V(q + \delta_a) + 2\delta,$$

where $\underline{\delta}$ is a vector of length m with every element equal to δ. So

$$|V(q + 2\underline{\delta}) - V(q + \delta_a)| > 2\delta = \|(q + 2\underline{\delta}) - (q + \delta_a)\|_\infty$$

and we have generated an expansion. Therefore a necessary condition for us to get a non-expansive map with respect to the L_∞ norm is that the partial derivatives $\frac{\partial V(q)}{\partial q_a}$ be non-negative.

Now, without loss of generality, assume that $q_1 = \max_a q_a$, so that $c_a = q_1 - q_a$ for each $a \in A$. Therefore we see that

$$V(q) = q_1 - \frac{\sum_a c_a h(c_a)}{\sum_a h(c_a)},$$

and so, for $m > 1$,

$$\frac{\partial V(q)}{\partial q_m} = \frac{\{h(c_m) - c_m h'(c_m)\} \sum_{a \geq 1} h(c_a) - h'(c_m) \sum_{a \geq 2} c_a h(c_a)}{\{\sum_{a \geq 1} h(c_a)\}^2}.$$

Choosing q such that $q_1 = q_2 = \ldots = q_{m-1} = q_m + c$, the numerator of this fraction becomes

$$\{h(c)\}^2 + (m-1)h(0)\{h(c) + ch'(c)\}.$$

Since we require this to be positive for any c and any value of m, this implies that

$$h(c) + ch'(c) \geq 0 \quad \forall c \geq 0. \tag{7.2}$$

Now, for general q,

$$\frac{\partial V(q)}{\partial q_1} = \frac{h(0)\{\sum_{a \geq 1} h(c_a)\} - \{\sum_{a \geq 2} c_a h'(c_a)\}\{\sum_{a \geq 1} h(c_a)\} + \{\sum_{a \geq 2} h'(c_a)\}\{\sum_{a \geq 2} c_a h(c_a)\}}{\{\sum_{a \geq 1} h(c_a)\}^2}.$$

Setting $m = 2k + 1$, and choosing q such that $c_2 = \cdots = c_{k+1} = c$ and $c_{k+2} = \cdots = c_{2k+1} = d$, the numerator becomes

$$\{h(0)\}^2 + kh(0)\{h(c) + h(d) - ch'(c) - dh'(d)\} + k^2(d-c)\{h'(c)h(d) - h'(d)h(c)\},$$

and again since we can take k as big as we want we require that

$$(d-c)\{h'(c)h(d) - h'(d)h(c)\}$$

be non-negative for all c and d, i.e.

$$d \geq c \quad \Rightarrow \quad \frac{h'(c)}{h(c)} \geq \frac{h'(d)}{h(d)},$$

or equivalently

$$\left(\frac{h'(c)}{h(c)}\right)' \leq 0. \tag{7.3}$$

Now since $h(c) > 0$ for all c (by assumption), we can write $h(c) = e^{-g(c)}$ for some function g. Further, since h is (not necessarily strictly) decreasing, we must have $g'(x) \geq 0$. Applying condition (7.2) we see that

$$1 - cg'(c) \geq 0,$$

and so $g'(c) \leq \frac{1}{c}$.

However, applying condition (7.3) we see that $g''(c) \geq 0$, and so $g'(c)$ is non-decreasing in c. We cannot have $g'(c) \equiv 0$, since then the condition $h(c) \to 0$ as $c \to \infty$ would fail, and so for sufficiently large c we find that $g'(c)$ is bounded away from 0. But this contradicts the fact that $g'(c) \leq \frac{1}{c}$, and so there is no such function h.

This shows that the method of McNamara *et al.* (1997) cannot result in a non-expansive map $Q \mapsto V$ with respect to the L_∞ norm, which is in some sense the natural norm to use when considering value iteration.

7.2 Non-expansive smooth best responses

Now consider smooth best responses (3.4) in the style of Harsanyi (1973), as used for stochastic fictitious play (Fudenberg and Kreps 1993), and the standard definition of the smooth best response dynamics (Hofbauer and Hopkins 2000). As in the previous section, we fix x, and write $q_a = Q(x,a)$, and $V(q)$ for the value of $V(x)$ resulting from the $Q(x,a)$. We see that

$$V(q) = \max_{\pi \in \Delta} \left\{ \sum_{a \in A} \pi(a) q_a + \tau v(\pi) \right\}.$$

Introduce a Lagrangian multiplier $\Lambda \in \mathbb{R}$ so that the maximisation becomes

$$\max_{\pi \in \mathbb{R}^{|A|}, \Lambda \in \mathbb{R}} \left\{ \sum_{a \in A} \pi(a) q_a + \tau v(\pi) + \Lambda \left(1 - \sum_{a \in A} \pi(a) \right) \right\},$$

assume that $v(\pi) = \sum_a \tilde{v}(\pi(a))$, and differentiate with respect to $\pi(a)$ to see that $\beta(q)(a)$ solves

$$q_a + \tau \nabla \tilde{v}(\beta(q)(a)) - \Lambda = 0.$$

The conditions on v specified in Section 1.1.2 imply that $\nabla \tilde{v}$ has a smooth inverse function h which is strictly decreasing and positive, and so

$$\beta(q)(a) = h\left(\frac{\Lambda - q_a}{\tau} \right),$$

where Λ is chosen so that $\sum_{a\in A}\beta(q)(a) = 1$.

We investigate how to choose h (and hence v) such that the resultant β is a non-expansive smooth best response, approaching the problem by trying to maximise $|V(q) - V(q')|$ for fixed arbitrary q', under the constraint that $\|q - q'\|_\infty \leq D$; if this maximal value is no greater than D then we are done.

If $\frac{\partial V(q)}{\partial q_a} \geq 0$ for all q then it is clear that $V(q)$ is maximised, subject to our constraints, by taking $q_a = q'(a) + D$ for each $a \in A$, at which point $V(q) = V(q')+D$. Similarly, $V(q)$ takes a minimum value of $V(q')-D$ when $q_a = q'(a)-D$ for all $a \in A$. So $\frac{\partial V(q)}{\partial q_a} \geq 0$ for all q is a sufficient condition to get a non-expansive map. As before, $\frac{\partial V(q)}{\partial q_a} \geq 0$ is also a necessary condition for the map $q \mapsto V(q)$ to be non-expansive.

Now, differentiating the condition $\sum_{b\in A} h(\tau^{-1}(\Lambda - q_b)) = 1$ with respect to q_a gives

$$\frac{\partial \Lambda}{\partial q_a} = \frac{h'(\tau^{-1}(\Lambda - q_a))}{\sum_{b\in A} h'(\tau^{-1}(\Lambda - q_b))}.$$

Since $V(q) = \sum_b h(\tau^{-1}(\Lambda - q_b))q_b$, it follows that

$$\begin{aligned}\frac{\partial V(q)}{\partial q_a} =& h(\tau^{-1}(\Lambda - q_a)) - \tau^{-1}h'(\tau^{-1}(\Lambda - q_a))q_a \\ &+ \tau^{-1}\frac{\partial \Lambda}{\partial q_a}\sum_{b\in A} h'(\tau^{-1}(\Lambda - q_b))q_b \\ =& h(\tau^{-1}(\Lambda - q_a)) + \tau^{-1}(\Lambda - q_a)h'(\tau^{-1}(\Lambda - q_a)) \\ &+ \tau^{-1}h'(\tau^{-1}(\Lambda - q_a))\left(\frac{\sum_{b\in A} h'(\tau^{-1}(\Lambda - q_b))q_b}{\sum_{b\in A} h'(\tau^{-1}(\Lambda - q_b))} - \Lambda\right).\end{aligned}$$

This is perhaps easier to read if we write $\tilde{c}_a = \tau^{-1}(\Lambda - q_a)$. Then

$$\frac{\partial V(q)}{\partial q_a} = h(\tilde{c}_a) + \tilde{c}_a h'(\tilde{c}_a) - h'(\tilde{c}_a)\frac{\sum_{b\in A} h'(\tilde{c}_b)\tilde{c}_b}{\sum_{b\in A} h'(\tilde{c}_b)}.$$

As we have shown, if this is non-negative for all q then we have a non-expansive map. Recalling that $h'(c) < 0$ for all c, sufficient conditions for this to be non-negative for any q are that $\Lambda \geq q_a$ for all a and $h(c) + ch'(c) \geq 0$ for all c.

Using this formulation, and taking $\tilde{v}(\pi(a)) = \log(\pi(a))$ so that $h(c) = c^{-1}$, it follows that this results in a non-expansive map. Therefore we have shown that

non-expansive distributions exist, even if they may not be particularly practical (Λ for this model is the solution of a degree $|A|$ polynomial).

Incidentally, taking $h(c) = e^{-c}$ gives the Boltzmann distribution. Notice that $h(c) + ch'(c) = (1 - c)e^{-c}$, which can clearly be negative. From this, it is easy to construct examples where the map $q \mapsto V(q)$ is expansive.

7.3 Conclusion

We have considered a modification of Bellman's equations for MDPs where players accept that they will not play optimally. This means that the possibility of experimenting with potentially disastrous outcomes will be factored into the value functions. However, it is not clear that there is a unique smooth best response to a fixed environment under these conditions (nor even that there is a unique solution to the modified Bellman equations (7.1)).

We showed that there is no form of smooth best response in the style of McNamara *et al.* (1997) which results in a non-expansive map with respect to the L_∞ norm for general action spaces.

On the other hand, we showed that using the formulation of Fudenberg and Kreps (1993) we can construct a non-expansive smooth best response. Using these non-expansive smooth best responses means that there is a unique solution of the modified Bellman equations (7.1) for a fixed Markovian environment (Littman 1996).

However, this is not sufficient to prove directly that there is a unique Nash distribution of Markov games; perhaps by using a suitable weighted maximum norm based on the rewards of the game (see Section 1.3.1) the extension to 2-player zero-sum stochastic games could be achieved.

Chapter 8

Further work

8.1 Actor–critic algorithms

Two-timescales stochastic approximation is a natural tool to study actor-critic learning algorithms, and asynchronous versions of the basic technique have been studied by Konda and Borkar (2000) and Borkar (2001). This suggests studying a simple extension of the actor–critic algorithm of Chapter 3 which is applicable in stochastic games. In this situation, the policy will be evaluated using either Q-learning (which is off-policy) or SARSA (the on-policy equivalent) then updating strategies using a smooth best response to the current Q values. However, the interaction between states introduces complications into the resultant differential equations, and a direct analysis using the techniques of Borkar (1998) has thus far proved elusive.

8.2 Q-learning

Similar issues arise when we try to generalise the individual Q-learning algorithm of Chapter 5 to stochastic games. However, for 2-player zero-sum stochastic games, since a normal form game has a unique Nash distribution, a generalisation of Szepesvári and Littman's result is highly promising, although interaction between

the players prevents an immediate application.

A further extension of our analysis of individual Q-learning would be to show that this algorithm will converge to Nash distribution values for any games in which the chain-recurrent set of the smooth best response dynamics consists solely of Nash distributions. Since we have shown that trajectories of the Q-learning ODE (5.2) are asymptotic pseudotrajectories of a semiflow closely related to the smooth best response dynamics, I believe this is more of a notational extension than anything more fundamental.

8.3 Multiple-timescales learning

Multiple-timescales learning, in both the actor–critic and Q-learning formulations, is a development allowing the algorithms to converge in games for which few (if any) previous learning algorithms (or evolutionary processes) are known to converge. It is crucial to the theoretical study of multiple-timescales stochastic processes that the fast processes can be shown to converge to unique fixed points for fixed values of the slow processes. There are two complementary areas of fruitful research here: studying methods of showing that the fast processes converge to unique fixed points, perhaps by extending the graphical analysis of Section 4.5, will assist in application of the current theory, while studying what will happen if there are multiple fixed points of the fast systems, probably by relating the stochastic approximation to the theory of singularly perturbed dynamical systems (Jones 1995), might allow the application of the techniques to a wider range of systems.

It is also clearly of interest to extend our multiple-timescales learning algorithms to stochastic games. Very similar issues arise here as have been discussed with respect to the previous extensions to stochastic games.

8.4 Discontinuous algorithms

Perhaps the area with greatest immediate promise is the study of discontinuous processes, such as the discontinuous actor–critic algorithm of Chapter 6. We have shown that these can be studied using ideas closely related to stochastic approximation, and even that a two-timescales approach is feasible. Although the *ad hoc* nature of the present approach to discontinuous systems is not wholly satisfactory, Delyon (1996) and Tadić (1998) have developed a theory of stochastic approximation with discontinuous dynamics, and Hofbauer and Sorin (2002) suggest that they are preparing a paper, along with Benaïm, which will extend Benaïm's methods of stochastic approximation to cover differential inclusions as well as the simple ordinary differential equations covered by Benaïm (1999). It is hoped that these general results can be extended to two timescales.

The work of Chapter 6 can be trivially extended to include a version of the actor–critic algorithm of Chapter 3 in which each player adjusts their strategy towards a smooth best response to the Q values, but with vanishing temperature parameters in the limit. Further, a modification of stochastic fictitious play in which the smooth best responses are calculated using vanishing temperatures can be studied in this framework too; it is clear that the beliefs will converge to a Nash equilibrium, but of more interest (and less obvious) is whether the strategies of the players must converge.

In Chapter 5 we showed that individual Q-learning with fixed temperature smooth best responses is closely related to the smooth best response dynamic. It is clearly of interest whether an asymptotically optimal version of this algorithm, where in the limit the players only play actions which maximise the Q value, can be shown to relate in some meaningful way to the best response dynamics. If this can be extended to stochastic games, it would justify a naive implementation of standard Q-learning in a multi-agent setting.

8.5 Convergence-rate analysis

A final area of research that has not been considered in this thesis is the rate of convergence of the algorithms. Although we have frequently shown that the limit sets of strategies will be contained in the set of Nash distributions, or the set of Nash equilibria, there is no indication of how long this convergence might take. Indeed the numerical results suggest that several million iterations is not sufficient for the strategies to be even particularly close to the equilibrium (for example Fig. 5.1). Konda and Tsitsiklis (2002) have studied some issues relating to the convergence rate of two-timescales stochastic approximation, but if algorithms are to be applied in real-world multi-agent settings it is important that progress be made to assist in the choosing of learning parameters for unknown environments.

Bibliography

Akin, E. and V. Losert (1984). Evolutionary dynamics of zero-sum games. *Journal of Mathematical Biology* 20, 231–258.

Aubin, J.-P. and A. Cellina (1984). *Differential Inclusions*. New York: Springer-Verlag.

Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (1995). Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *36th Annual Symposium on Foundations of Computer Science*, pp. 322–331. Los Alamitos, CA: IEEE Press.

Baños, A. (1968). On pseudo-games. *The Annals of Mathematical Statistics* 39, 1932–1945.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America* 38, 716–719.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Le Séminaire de Probabilités*, Volume 1709 of *Lecture Notes in Mathematics*. New York: Springer–Verlag.

Benaïm, M. and M. W. Hirsch (1999). Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behaviour* 29, 36–72.

Benaïm, M., J. Hofbauer, and S. Sorin (2003). Stochastic approximations and differential inclusions. Available at `http://www.unine.ch/math/`

personnel/equipes/benaim/benaim_pers/bhs.pdf.

Berger, U. (2003). Fictitious play in $2 \times m$ games. Ordinariat VW 5 Discussion Paper 0301, Vienna University of Economics. Presented at the 2003 Game Theory Festival at Stonybrook.

Bertsekas, D. P. and J. N. Tsitsiklis (1996). *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

Börgers, T. and R. Sarin (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* **77**, 1–14.

Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems and Control Letters* **29**, 291–294.

Borkar, V. S. (1998). Ashyncronous stochastic approximations. *SIAM Journal of Control and Optimization* **36**, 840–851.

Borkar, V. S. (2001). Reinforcement learning in Markovian evolutionary games. Available at http://www.tcs.tifr.res.in/~borkar/game.ps.

Bowling, M. (2000). Convergence problems of general-sum multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 89–94. San Francisco, CA: Morgan Kaufmann.

Bowling, M. and M. Veloso (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence* **136**, 215–250.

Boyan, J. A. and M. L. Littman (1994). Packet routing in dynamically changing networks: A reinforcement learning approach. In J. D. Cowan, G. Tesauro, and J. Alspector (Eds.), *Advances in Neural Information Processing Systems*, Volume 6, pp. 671–678. San Francisco, CA: Morgan Kaufmann.

Brandière, O. (1998a). The dynamic system method and the traps. *Advances in Applied Probability* **30**, 137–151.

Brandière, O. (1998b). Some pathological traps for stochastic approximation. *SIAM Journal of Control and Optimization* **36**, 1293–1314.

Brown, G. W. (1951). Iterative solution of games by fictitious play. In T. C. Koopmans (Ed.), *Activity Analysis of Production and Allocation*, pp. 374-376. New York: John Wiley & Sons, Inc.

Bush, R. R. and F. Mosteller (1951). A mathematical model for simple learning. *Psychological Review* 58, 313-323.

Chen, H.-F. and Y.-M. Zhu (1986). Stochastic approximation procedures with randomly varying truncations. *Scientia Sinica, Series A* 29, 914-926.

Claus, C. and C. Boutilier (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98) and the Tenth Conference on Innovative Applications of Artificial Intelligence (IAAI-98)*, pp. 746–752. Menlo Park, CA: AAAI Press.

Cowan, S. (1992). *Dynamical Systems Arising from Game Theory*. Ph. D. thesis, University of California, Berkeley.

Crites, R. H. and A. G. Barto (1995). An actor/critic algorithm that is equivalent to Q-learning. In G. Tesauro, D. Touretzky, and T. Leen (Eds.), *Advances in Neural Information Processing Systems*, Volume 7, pp. 401–408. Cambridge, MA: MIT Press.

Crites, R. H. and A. G. Barto (1998). Elevator group control using multiple reinforcement learning agents. *Machine Learning* 33, 235-262.

Delyon, B. (1996). General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control* 41, 1245-1255.

Filar, J. and K. Vrieze (1997). *Competitive Markov Decision Processes*. New York: Springer-Verlag.

Foster, D. P. and R. V. Vohra (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior* 21, 40-55.

BIBLIOGRAPHY

Foster, D. P. and R. V. Vohra (1998). Asymptotic calibration. *Biometrika* **85**, 379–390.

Foster, D. P. and R. V. Vohra (1999). Regret in the on-line decision problem. *Games and Economic Behavior* **29**, 7–35.

Fudenberg, D. and D. M. Kreps (1993). Learning mixed equilibria. *Games and Economic Behavior* **5**, 320–367.

Fudenberg, D. and D. K. Levine (1998). *The Theory of Learning in Games.* Cambridge, MA: MIT Press.

Fudenberg, D. and D. K. Levine (1999). Conditional universal consistency. *Games and Economic Behaviour* **29**, 104–130.

Fudenberg, D. and J. Tirole (1991). *Game Theory.* Cambridge, MA: MIT Press.

Gaunersdorfer, A. and J. Hofbauer (1995). Fictitious play, Shapley polygons, and the replicator equation. *Games and Economic Behavior* **11**, 279–303.

Gilboa, I. and A. Matsui (1991). Social stability and equilibrium. *Econometrica* **59**, 859–867.

Govindan, S., P. J. Reny, and A. J. Robson (2003). A short proof of Harsanyi's purification theorem. *Games and Economic Behavior* **45**, 369–374.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Drescher, A. W. Tucker, and P. Wolfe (Eds.), *Contributions to the theory of games, Vol. 3*, Volume 39 of *Annals of Mathematical Studies*, pp. 97–139. Princeton University Press.

Harsanyi, J. C. (1973). Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory* **2**, 1–23.

Harsanyi, J. C. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games.* Cambridge, MA: MIT Press.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 1127–1150.

Hart, S. and A. Mas-Colell (2001a). A general class of adaptive strategies. *Journal of Economic Theory* 98, 26–54.

Hart, S. and A. Mas-Colell (2001b). A reinforcement procedure leading to correlated equilibrium. In W. N. G. Debreu and W. Trockel (Eds.), *Economic Essays: A Festschrift for Werner Hildenbrand*, pp. 181–200. New York: Springer–Verlag.

Hart, S. and A. Mas-Colell (2003). Uncoupled dynamics cannot lead to Nash equilibrium. *American Economic Review* **23**, 1830–1836.

Hofbauer, J. (1995). Stability for the best response dynamics. Technical report, Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A–1090 Vienna, Austria.

Hofbauer, J. (1996). Evolutionary dynamics for bimatrix games: A Hamiltonian system? *Journal of Mathematical Biology* **34**, 675–688.

Hofbauer, J. (2000). From Nash and Brown to Maynard Smith: Equilibria, dynamics, and ESS. *Selection* **1**, 81–88.

Hofbauer, J. and E. Hopkins (2000). Learning in perturbed asymmetric games. Available at `http://www.ed.ac.uk/~ehk/perturb.pdf`.

Hofbauer, J. and W. H. Sandholm (2002). On the global convergence of stochastic fictitious play. *Econometrica* **70**, 2265–2294.

Hofbauer, J. and K. Sigmund (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Hofbauer, J. and S. Sorin (2002). Best response dynamics for continuous zero-sum games. Technical Report 2002-028, Laboratoire d'econometric, Ecole Polytechnique.

BIBLIOGRAPHY

Hofbauer, J. and J. W. Weibull (1996). Evolutionary selection against dominated strategies. *Journal of Economic Theory* **71**, 558–573.

Hopkins, E. (1999). A note on best response dynamics. *Games and Economic Behavior* **29**, 138–150.

Hu, J. and M. P. Wellman (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 242–250. San Francisco, CA: Morgan Kaufmann.

Jaakkola, T., M. I. Jordan, and S. Singh (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation* **6**, 1185–1201.

John, G. H. (1994). When the best move isn't optimal: Q-learning with exploration. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-94)*. Menlo Park, CA: AAAI Press.

Jones, C. K. R. T. (1995). Geometric singular perturbation theory. In *Dynamical Systems*, Volume 1609 of *Lecture Notes in Mathematics*, pp. 44–118. Berlin: Springer-Verlag.

Jordan, J. S. (1993). Three problems in learning mixed strategy equilibria. *Games and Economic Behavior* **5**, 368–386.

Kalai, E. and E. Lehrer (1993a). Rational learning leads to Nash equilibrium. *Econometrica* **61**, 1019–1045.

Kalai, E. and E. Lehrer (1993b). Subjective equilibrium in repeated games. *Econometrica* **61**, 1231–1240.

Koller, D. and B. Milch (2003). Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior* **45**, 181–221.

Konda, V. R. and V. S. Borkar (2000). Actor-critic-type learning algorithms for Markov decision process. *SIAM Journal on Control and Optimization* **38**,

94–123.

Konda, V. R. and J. N. Tsitsiklis (2002). Convergence rate of two-time-scale stochastic approximation. Submitted to *Annals of Applied Probability*, March 2002.

Kushner, H. J. and D. S. Clark (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer–Verlag.

Kushner, H. J. and G. G. Yin (1997). *Stochastic Approximation Algorithms and Applications*. New York: Springer–Verlag.

Leslie, D. S. and E. J. Collins (2003). Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Annals of Applied Probability* **13**, 1231–1251.

Littman, M. L. (1996). *Algorithms for Sequential Decision Making*. Ph. D. thesis, Department of Computer Science, Brown University, Providence, RI.

Littman, M. L. and P. Stone (2001). Implicit negotiation in repeated games. In J.-J. Ch. Meyer and M. Tambe (Eds.), *Intelligent Agents VIII: Agent Theories, Architectures, and Languages*, Volume 2333 of *Lecture Notes in Computer Science*, pp. 393–404. New York: Springer–Verlag.

Littman, M. L. and C. Szepesvári (1996). A generalized reinforcement-learning model: Convergence and applications. In L. Saitta (Ed.), *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 310–318. San Francisco, CA: Morgan Kaufmann.

Littman, M. L., M. Kearns, and S. Singh (2001). An efficient, exact algorithm for solving tree-structured graphical games. In T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*, Volume 14. Cambridge, MA: MIT Press.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control* **22**, 551–575.

BIBLIOGRAPHY

Maynard Smith, J. (1982). *Evolution and the Theory of Games.* Cambridge University Press.

McNamara, J. M., J. N. Webb, E. J. Collins, T. Székely, and A. I. Houston (1997). A general technique for computing evolutionarily stable strategies based on errors in decision-making. *Jounal of Theoretical Biology* **189**, 211–225.

Megiddo, N. (1980). On repeated games with incomplete information played by non-Bayesian players. *International Journal of Game Theory* **9**, 157–167.

Milgrom, P. and J. Roberts (1991). Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior* **3**, 82–100.

Monderer, D. and L. S. Shapley (1996). Fictitious play property for games with identical interests. *Journal of Economic Theory* **68**, 258–265.

Narendra, K. S. and M. A. L. Thathachar (1989). *Learning Automata: An Introduction.* Englewood Cliffs, NJ: Prentice-Hall.

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America* **36**, 48–49.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics* **54**, 286–295.

Pemantle, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *Annals of Probability* **18**, 698–712.

Plank, M. (1997). Some qualitative differences between the replicator dynamics of two player and n player games. *Nonlinear Analysis, Theory, Methods & Applications* **30**, 1411–1417.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Chichester: John Wiley & Sons, Inc.

Ritzberger, K. and J. W. Weibull (1995). Evolutionary selection in normal form games. *Econometrica* **63**, 1371–1399.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **22**, 400–407.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics* **54**, 296–301.

Ross, S. M. (1982). *Introduction to stochastic dynamic programming*. New York: Academic Press.

Rummery, G. A. (1995). *Problem Solving with Reinforcement Learning*. Ph. D. thesis, Cambridge University.

Sato, Y. and J. P. Crutchfield (2002). Coupled replicator equations for the dynamics of learning in multiagent systems. Technical Report 02-04-017, Santa Fe Institute.

Schraudolph, N. N., P. Dayan, and T. J. Sejnowski (1994). Temporal difference learning of position evaluation in the game of go. In J. D. Cowan, G. Tesauro, and J. Alspector (Eds.), *Advances in Neural Information Processing*, Volume 6, pp. 817–824. San Francisco, CA: Morgan Kaufmann.

Schuster, K. P. and K. Sigmund (1981). Coyness, philandering and stable strategies. *Animal Behavior* **29**, 186–192.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America* **39**, 1095–1100.

Shapley, L. S. (1964). Some topics in two person games. In M. Drescher, L. S. Shapley, and A. W. Tucker (Eds.), *Advances in Game Theory*. Princeton University Press.

Singh, S. and D. Bertsekas (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, Volume 9, pp. 974. Cambridge, MA: MIT Press.

BIBLIOGRAPHY

Singh, S. and R. S. Sutton (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning* **23**, 123–158.

Singh, S., T. Jaakkola, M. L. Littman, and C. Szepesvari (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning* **38**, 287–308.

Singh, S., M. Kearns, and Y. Mansour (2000). Nash convergence of gradient dynamics in general-sum games. In *Uncertainty in Artificial Intelligence (UAI): Proceedings of the 16th Conference (UAI'00)*. San Francisco, CA: Morgan Kaufmann.

Stone, P. (2000). *Layered Learning in Multiagent Systems: A Winning Approach to Robotic Soccer*. Cambridge, MA: MIT Press.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning* **3**, 9–44.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Szepesvári, C. and M. L. Littman (1999). A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Computation* **11**, 2017–2059.

Tadić, V. (1998). Stochastic approximation with random truncations, state-dependent noise and discontinuous dynamics. *Stochastics and Stochastics Reports* **64**, 283–326.

Tesauro, G. (1994). TD-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation* **6**, 215–219.

Thorndike, E. (1898). Some experiments on animal intelligence. *Science* **7**, 818–824.

Van der Genugten, B. (2000). A weakened form of fictitious play in two-person zero-sum games. *International Game Theory Review* **2**, 307–328.

Von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* **100**, 295–320.

Von Neumann, J. and O. Morgenstern (1953). *Theory of games and economic behavior*. Princeton University Press.

Vrieze, O. J. and S. H. Tijs (1982). Fictitious play applied to sequences of games and discounted stochastic games. *International Journal of Game Theory* **11**, 71–85.

Watkins, C. J. C. H. and P. Dayan (1992). *Q*-learning. *Machine Learning* **8**, 279–292.

Watkins, C. J. H. (1989). *Learning from Delayed Rewards*. Ph. D. thesis, Cambridge University.

Williams, R. J. and L. C. Baird (1993). Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems. Technical Report NU-CCS-93-11, College of Computer Science, Northeastern University.